# Momentos geométricos y machine learning aplicados al estudio de datos oftalmológicos

**Trabajo de Fin de Grado**
**Curso 2019–2020**


**Doble Grado en Ingeniería Informática y Matemáticas**
**Facultad de Ciencias Matemáticas**
**Universidad Complutense de Madrid**

**Autor**
**Carlos Moreno Morera**

**Director**
**Carlos Gregorio Rodríguez**

**6 de octubre de 2020**

# Momentos geométricos y machine learning aplicados al estudio de datos oftalmológicos

**Trabajo de Fin de Grado en Matemáticas**
**Departamento de Sistemas Informáticos y Computación**


**Autor**
**Carlos Moreno Morera**

**Director**
**Carlos Gregorio Rodríguez**


**Convocatoria:** *Septiembre 2020*


**Doble Grado en Ingeniería Informática y Matemáticas**
**Facultad de Ciencias Matemáticas**
**Universidad Complutense de Madrid**


**6 de octubre de 2020**

*A mis padres, por siempre apoyarme y confiar en que podría con todos los retos.*

# Resumen


## Momentos geométricos y machine learning aplicados al estudio de datos oftalmológicos

El glaucoma es la primera causa de ceguera irreversible en el ámbito mundial. En su desarrollo, aumenta la excavación presente en la papila óptica de la retina. Esta depresión puede ser examinada mediante la tomografía de coherencia óptica y parametrizada como una superficie regular mediante la utilización de momentos geométricos invariantes a transformaciones de semejanza. Estas entidades matemáticas se pueden agrupar para formar un conjunto de datos de pacientes glaucomatosos y personas sanas que puede ser estudiado mediante técnicas estadísticas básicas y de aprendizaje automático. En este trabajo, se analiza dicho conjunto para evaluar el potencial de los momentos geométricos y de los métodos de machine learning en la investigación con datos oftalmológicos. Se estudian propiedades y características de los invariantes que permiten mejorar su entendimiento y aportar información sobre la enfermedad. Algunas de las cualidades que se examinan son la consistencia, la dispersión, la correlación entre los momentos, cómo se estructuran en el espacio, la relevancia de cada invariante en el diagnóstico del glaucoma y las muestras del entorno de los distintos elementos del conjunto en el espacio. Para abordar estas cuestiones, se hará uso de técnicas estadísticas como los estadísticos descriptivos más comunes, el coeficiente de correlación de Pearson y el error cuadrático medio; y de métodos de machine learning como los algoritmos de clustering, el análisis de componentes principales y los árboles de decisión. De esta forma, se lleva a cabo una primera prueba de concepto, sobre los datos en bruto, con la que se investigan las distintas cuestiones que pueden abordarse y con la que se plantean nuevos temas y preguntas, de manera que se alcancen conclusiones que aporten información sobre esta neuropatía óptica desde el campo de las matemáticas a la rama de la oftalmología.



## Palabras clave

Aprendizaje automático, momentos geométricos invariantes, glaucoma, tomografía de coherencia óptica, K-Medias, DBSCAN, análisis de componentes principales, árboles de decisión, importancia de Gini, estadísticos descriptivos

# Abstract


## Surface moments and machine learning applied to the study of ophthalmological data

Glaucoma is the leading cause of irreversible blindness worldwide. In its development, it increases the excavation present in the optic disc of the retina. This depression can be examined by optical coherence tomography and parameterised as a regular surface by using three-dimensional surface moments invariant under similarity transformations. Thanks to these mathematical entities, we have a data set of glaucomatous patients and healthy people that can be studied by means of basic statistical techniques and machine learning methods. In this work, this set is analysed to evaluate the potential of surface moments and machine learning methods in ophthalmological data research. Properties and characteristics of the invariants are studied to improve their understanding and provide information about the disease. Some of the qualities examined are consistency, dispersion, correlation between the moments, how they are structured in space, the relevance of each invariant in the diagnosis of glaucoma and the environmental samples of the different elements of the set in space. To address these issues, we will use statistical techniques such as the most common descriptive statistics, the Pearson's correlation coefficient and the mean square error; and machine learning methods such as clustering algorithms, principal component analysis and decision trees. In this way, a first proof of concept is carried out, on this raw data, with which the different issues that can be addressed are researched and with which new topics and questions are raised, so that conclusions are reached that provide information on this optical neuropathy from the field of mathematics to the branch of ophthalmology.



## Keywords

Machine learning, surface moment invariants, glaucoma, optical coherence tomography, K-Mans, DBSCAN, principal component analysis, decision trees, Gini importance, descriptive statistics

# Índice

# Índice de figuras

# Índice de tablas

Capítulo 1

# Introducción

*"La investigación de las enfermedades ha avanzado tanto que es cada vez más difícil encontrar a alguien que esté completamente sano."*
— Aldous Huxley

En este capítulo se explican los motivos que incentivan el desarrollo de este trabajo (véase la sección 1.1) y los objetivos que se proponen para ser alcanzados (véase la sección 1.2) durante su elaboración. A continuación, se explica la estructura de esta memoria y el orden que siguen los distintos capítulos (véase la sección 1.3). Por último, se presentan las herramientas software utilizadas a lo largo de todo el estudio (véase la sección 1.4).

## 1.1. Motivación

El aumento de capacidad de procesamiento de los computadores ha hecho posible la ejecución y el avance de las técnicas de aprendizaje automático, ya que, por lo general, los métodos de machine learning poseen un coste computacional elevado cuando se tratan cantidades significativas de datos. Este desarrollo ha permitido aplicar estas técnicas a campos de investigación distintos al de las matemáticas y las ciencias de la computación con resultados con una relevancia bastante significativa. Uno de estos campos en el que es posible encontrar numerosos estudios multidisciplinarios es la medicina. En este trabajo se pretende aplicar técnicas de aprendizaje automático a datos médicos, concretamente pertenecientes a la rama de la oftalmología.

El glaucoma es una de las mayores causas de ceguera en el ámbito mundial y la primera causa de ceguera irreversible. Además, suele ser asintomática hasta que alcanza fases avanzadas, por lo que el diagnóstico precoz se ha convertido en un problema sanitario de gran relevancia en la actualidad. Se caracteriza por un aumento de la excavación presente en la papila óptica (o punto ciego) de la retina. Una de las técnicas para observar la cúpula (que es el nombre que posee dicha zona del punto ciego) es la tomografía de coherencia óptica. Gracias a ella puede examinarse esta depresión, la cual, desde un punto de vista matemático, puede caracterizarse como una superficie regular. Esta interpretación que otorga una definición formal a un concepto oftalmológico, puede dar lugar a numerosos estudios multidisciplinarios que aprovechen el amplio conocimiento de las superficies regulares, así como del glaucoma. Precisamente, como se trata de un objeto geométrico, es posible parametrizar la excavación de la papila óptica mediante el uso de momentos geométricos invariantes a transformaciones de semejanza. Gracias a estas entidades matemáticas pue-

de llevarse a cabo un estudio de esta enfermedad desde el punto de vista del aprendizaje automático para abordar distintas cuestiones a partir del conjunto de datos, y así alcanzar conclusiones que aporten más información sobre esta neuropatía óptica desde el campo de las matemáticas a la rama de la oftalmología.

## 1.2. Objetivos

Ante un conjunto de datos que no ha sido estudiado previamente, el principal objetivo es analizar el potencial que podrían tener para seguir desarrollando el proyecto de análisis con una población con un tamaño mayor. Para ello, dado que, como se verá, los momentos geométricos invariantes utilizados son poco explicables, será necesario extraer conclusiones de sus propiedades y características y, así, poder ir mejorando la comprensión que se tiene de ellos. De esta manera, con este trabajo se pretende llevar a cabo una primera prueba de concepto con la que se pueda tantear las distintas cuestiones que pueden abordarse y plantear nuevos temas que no surgirían en una primera instancia. Por esta razón, se tratará de resolver tanto simples preguntas como la correlación existente entre cada par de invariantes, como preguntas que requieran técnicas más complejas como la relevancia que tiene cada uno de ellos a la hora de diagnosticar la enfermedad.

En lo que se refiere a las cuestiones más "sencillas" (es decir, que requieran tan solo de técnicas de estadística básica) se pretende conocer la consistencia de los momentos, sus estadísticos descriptivos más comunes, su dispersión y la correlación presente entre ellos. Con invariantes consistentes nos referimos a que, en caso de llevar a cabo la prueba a un mismo paciente, se obtengan resultados similares y, como cabría esperar, con un considerable ratio de diferenciación respecto del resto de individuos. Mientras que preguntas complejas que pueden ser planteadas son la estructura en el espacio con la que se organizan los distintos elementos del sistema de datos, la relevancia que tiene cada uno de los momentos geométricos a la hora de diagnosticar la enfermedad o si el entorno de los pacientes glaucomatosos son individuos sanos o enfermos y viceversa.

Además de conocer cuánto partido es posible sacarle a los momentos geométricos, otro objetivo es el de evaluar la eficacia de las distintas técnicas de machine learning sobre los datos oftalmológicos. De esta manera, resulta interesante conocer si es posible aplicar alguno de estos métodos para automatizar alguna tarea (como el trabajo de diagnóstico) o para descubrir información sobre la enfermedad que pueda ser de utilidad.

## 1.3. Estructura del trabajo

Siguiendo los objetivos del trabajo explicados en la sección anterior, esta memoria se estructura en cinco capítulos diferentes. El primero de ellos, como se ha visto, se trata de la introducción al estudio realizado, donde se pretende conformar una primera percepción de lo que se explicará a lo largo del resto del trabajo. En el capítulo 2 (titulado "Conociendo el origen de los datos") se presenta el conocimiento previo que se posee del conjunto de datos. De esta manera, se lleva a cabo una breve introducción al glaucoma, la enfermedad que se estudia, para poder llegar a comprender su sintomatología y cómo se desarrolla en el ojo humano. A continuación, se presenta la técnica de tomografía de coherencia óptica, con la que es posible diagnosticar esta neuropatía óptica. Resulta importante conocer ciertos detalles de este método para comprender cómo se estructura los datos extraídos del paciente que se analizarán más adelante. Por último, como la tomografía obtiene una "superficie", este capítulo finaliza definiendo formalmente los momentos geométricos invariantes que la

describirán y con los que se tratará a lo largo del resto del trabajo.

El siguiente capítulo (el capítulo 3: "Presentación del conjunto de datos") es una primera aproximación al sistema de datos con el que se cuenta. En él se introducen las variables que describen el conjunto de datos y se utilizan técnicas de estadística básica (como el error cuadrático medio, la media, la mediana, la moda y el coeficiente de correlación de Pearson) para conocer las propiedades y características más fácilmente detectables que presenta el sistema. Entre las cualidades estudiadas se encuentran la consistencia de los momentos geométricos, su dispersión y la correlación que posee entre ellos. Estas propiedades deberán ser tenidas en cuenta a la hora de aplicar técnicas de aprendizaje automático, ya que podrían influir en su resultado.

En cuanto al capítulo 4 (Análisis de datos con técnicas de machine learning), en él se lleva a cabo el estudio de los datos mediante el uso de técnicas de aprendizaje automático. Se comienza con un análisis preliminar de los momentos utilizando algoritmos de clustering (K-Medias y DBSCAN) para comprobar si se los elementos del sistema se organizan en el espacio de acuerdo con su categoría sanitaria (pacientes que padecen glaucoma y individuos del grupo de control) y si es posible agruparlos de acuerdo con ella. A continuación, para comprender la estructura del conjunto de datos, se lleva a cabo un análisis de componentes principales. El capítulo finaliza con la propuesta de un modelo de clasificación que hace uso de los árboles de decisión, así como de la determinación de la relevancia de cada momento a la hora de diagnosticar la enfermedad.

Finalmente, en el capítulo 5 (Conclusiones y trabajo futuro), se exponen las conclusiones alcanzadas a lo largo de todo el estudio y las posibles mejoras y extensiones que pueden ser llevadas a cabo con el propósito de continuar trabajando sobre este tema de investigación.

## 1.4. Herramientas software utilizadas

A lo largo del trabajo se han aplicado diversas técnicas estadísticas y de aprendizaje automático que requieren del uso de un computador para su evaluación. Para su desarrollo se ha utilizado el lenguaje de programación *Python* en el entorno interactivo *Jupyter Notebook*. La ventaja que tiene este lenguaje sobre otros compilados es su particularidad de ser interpretado. Gracias a esta característica posee un modo interactivo en el que las instrucciones pueden escribirse una a una, o en grupos, y obtener la visualización del resultado de la evaluación de cada una de manera inmediata. Esto da la posibilidad de probar porciones de código antes de integrarlo como parte de un programa final.

En concreto, principalmente se han aprovechado las funcionalidades de las librerías *sklearn* (que proporciona la implementación de las técnicas de aprendizaje automático utilizadas a lo largo de este trabajo), *pandas* (que facilita la gestión de los conjuntos de datos como *DataFrames* y el cálculo de los estadísticos descriptivos más comunes de estas estructuras) y *numpy* (que facilita la gestión de vectores y matrices, así como implementa el cálculo de la matriz de correlación dado un conjunto de datos). Estos recursos, han evitado la implementación de las distintas técnicas que se han requerido, a excepción del cálculo de la importancia de Gini (véase la sección 4.3.2), el cual se ha programado haciendo uso de la información que almacena el módulo *tree* de la primera librería mencionada.

Capítulo 2

# Conociendo el origen de los datos

*"¡Datos! ¡Datos! ¡Datos! - exclamó con impaciencia - ¡No puedo fabricar ladrillos si no tengo arcilla!"*
— El misterio de Copper Beeches - Arthur Conan Doyle (1892)

Para poder llegar a comprender el análisis del conjunto de datos y extraer conclusiones con una relevancia significativa que permitan aprovechar las distintas técnicas de aprendizaje automático en este ámbito de la oftalmología, es necesario entender el origen de estos datos: cómo han sido extraídos y cómo han sido tratados previamente. Precisamente, la extracción y el tratamiento de los datos son el eje central de desarrollo de este capítulo.

Para explicar cómo han sido extraídos los datos hasta poder tratarlos, en la sección 2.1, se introduce una sencilla definición de glaucoma y su sintomatología. Previamente se lleva a cabo una breve y necesaria explicación sobre la anatomía del ojo humano. Específicamente, de la zona de la retina donde se sitúa el punto ciego, ya que examinando dicho área se logra diagnosticar esta enfermedad. Asimismo, se expone un método de diagnóstico muy común para esta y otras enfermedades: la tomografía de coherencia óptica, debido a que esta técnica nos proporcionará los datos preliminares que, tras ser tratados, se analizan en los capítulos siguientes y de los que se extraen las respectivas conclusiones de este estudio que pueden consultarse en el capítulo 5.

Dado que el formato de los datos extraídos de una tomografía de coherencia óptica no pueden ser estudiados con las técnicas de aprendizaje automático (ya que genera una estructura de datos compleja para este tipo de algoritmos), es necesario transformarlos a un formato que sí nos permita indagar en ellos. Por esta razón, en la sección 2.2, se explican los momentos geométricos invariantes que parametrizarán la superficie obtenida mediante la tomografía y las propiedades de estos.

Tras comprender este proceso desde que se extrae la información requerida del ojo del paciente hasta que, tratándolos, se convierten en un conjunto de tuplas analizables con las técnicas de aprendizaje automático, se podrá proceder al análisis de datos.

## 2.1. Glaucoma y Tomografía de Coherencia Óptica

El ojo humano (Netter, 2011) es un órgano que puede ser descrito, a modo de simplificación, como una esfera hueca llena de un gel trasparente llamado humor vítreo. Este líquido completa el espacio comprendido entre la retina y el cristalino (véase la figura 2.1). En la zona anterior del ojo, podemos encontrar un volumen entre la córnea y el iris conocido como cámara anterior. A continuación, entre el iris y el cristalino, se ubica la cámara

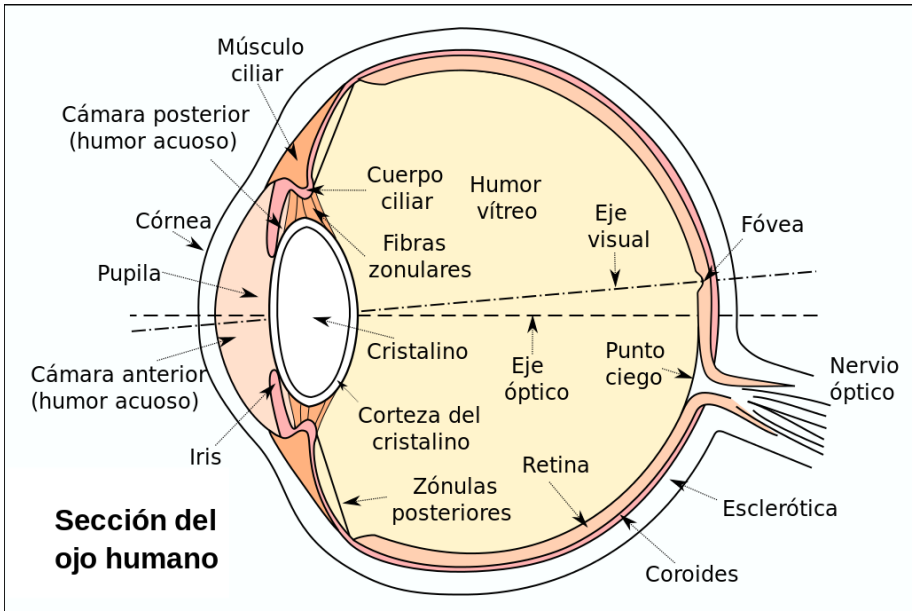


Figura 2.1: Esquema de la sección del ojo humano

posterior. Estos dos espacios contienen el líquido denominado humor acuoso que, junto con el humor vítreo anteriormente mencionado, conforman el conjunto de líquidos intraoculares. El nivel de la presión que ejercen estos líquidos contra la pared del ojo (denominada presión intraocular) es absolutamente trascendental para el correcto funcionamiento de este órgano. De hecho, un aumento patológico de la presión intraocular por un drenaje insuficiente del humor acuoso es el principal factor de riesgo para que se origine el glaucoma (Pavan-Langston, 2008; Noguera, 2012), una neuropatía óptica que puede provocar un deterioro significativo de la capacidad visual y es una de las mayores causas de ceguera en el ámbito mundial y se considera que, en la actualidad, es la primera causa de ceguera irreversible a nivel mundial (Grupo, 2011).

Como se puede observar en la figura 2.1, en el centro de la retina se encuentra el punto ciego, también conocido como papila óptica. Esta zona circular, por la que aparecen los axones de un tipo de neuronas llamadas células ganglionares de la retina (Tortora y Derrickson, 2013), se caracteriza por la carencia de conos y bastones, lo que imposibilita la existencia de sensibilidad ante estímulos luminosos en dicha superficie (razón por la cual se denomina punto ciego). Dentro de la papila se encuentra una excavación fisiológica llamada cúpula, cuyas dimensiones diametrales en comparación con las del punto ciego nos permiten también sospechar el diagnóstico de glaucoma (Michelessi et al., 2015).

Según Grupo (2011), a pesar de lo que suele pensarse, solo se puede hablar de glaucoma cuando existe daño del nervio óptico (si solo se observa una elevación aislada de la presión intraocular se debe considerar hipertensión ocular exclusivamente). Esta enfermedad no se define como aumento de presión intraocular. Esta confusión se debe a que, como se ha mencionado, esta neuropatía está frecuentemente relacionada con la elevación de la presión intraocular. De hecho puede padecerse glaucoma y haber tensión ocular normal.

El daño glaucomatoso supone una pérdida de fibras nerviosas que se observa estructuralmente como un aumento de la excavación papilar, y funcionalmente, como alteraciones del campo visual de inicio periférico. La agudeza visual se conserva hasta fases avanzadas, por lo que es poco sintomática. Por este motivo, la detección precoz de pacientes glaucomatosos se convierte en un problema sanitario importante. Concretamente, es preciso realizar

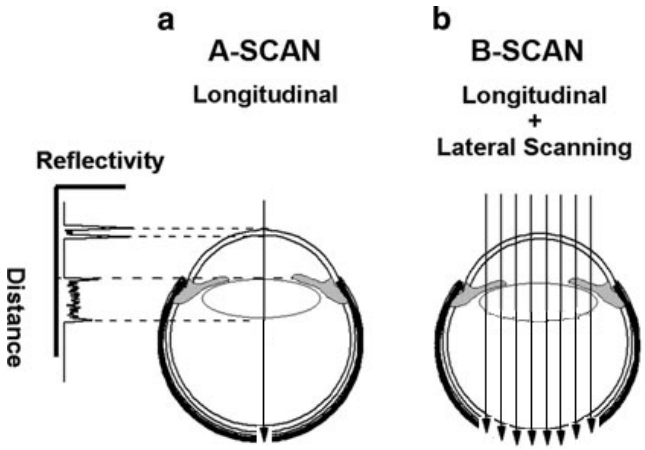


Figura 2.2: Esquema de los cortes trasversales y axiales de la OCT

un diagnóstico precoz en todo individuo mayor de 40 años o que presente factores de riesgo (como antecedentes familiares, miopía, síndrome de apnea del sueño y paquimetría corneal baja o córnea fina).

El daño anatómico del nervio óptico se valora mediante la examinación visual de la excavación papilar o, recientemente, mediante análisis computarizados del grosor de la capa de fibras nerviosas de la retina (OCT, GDX o HRT, entre otros). Una de las técnicas más comunes utilizadas para el diagnóstico del glaucoma es la tomografía de coherencia óptica (Huang et al., 1991). Esta tecnología (cuyas siglas son OCT por su denominación inglesa *Optical Coherence Tomography*) es una prueba no invasiva del fondo del ojo que obtiene imágenes de las capas de la retina con una gran resolución. Gracias a ellas, es posible evaluar la papila, cuantificar el espesor de la capa de fibras nerviosas y estudiar el cociente entre el diámetro de la cúpula con el de la papila óptica para diagnosticar esta enfermedad.

Sin duda, donde más revolucionaria ha resultado la OCT es en el estudio de las enfermedades retinianas, aunque, según Grupo (2011), algunos estudios sugieren que los pacientes de Parkinson y Alzheimer presentan una reducción significativa en la capa de fibras nerviosas y en la capa de células ganglionares, pudiendo resultar útil esta técnica en el manejo de esos pacientes. Una enfermedad neurológica en la que la OCT ha demostrado una indiscutible utilidad es la esclerosis múltiple, debido a que los pacientes presentan reducción significativa en el espesor de la capa de fibras retinianas. También puede ser aprovechada para la detección y seguimiento de otras neuropatías como lo son la neuritis o drusas del nervio.

Durante la tomografía, se lleva a cabo otra prueba: la oftalmoscopia con láser de barrido (Roorda et al., 2002; Mainster et al., 1982), cuyas siglas son SLO por su denominación inglesa *Scanning Laser Ophthalmoscopy*. La imagen SLO, resultante de la oftalmoscopia, es una imagen bidimensional del fondo del ojo, es decir, de la retina. Esta técnica se utiliza para centrar correctamente la OCT sobre la papila óptica. El resultado final de una tomografía de coherencia óptica consta de un número natural de *B-Scans* comprendido entre 1 y 241, cada uno de los cuales se construye con una cantidad de *A-scans* entre 384 y 1536. Estos últimos contienen información sobre las dimensiones y la localización espacial de estructuras situadas dentro del ojo (véase la figura 2.2). Se conocen como exploraciones de profundidad axial. Al combinar una serie de A-Scans se obtiene un tomógrafo de corte transversal (B-scan), gracias al cual se consigue una imagen bidimensional de las distintas

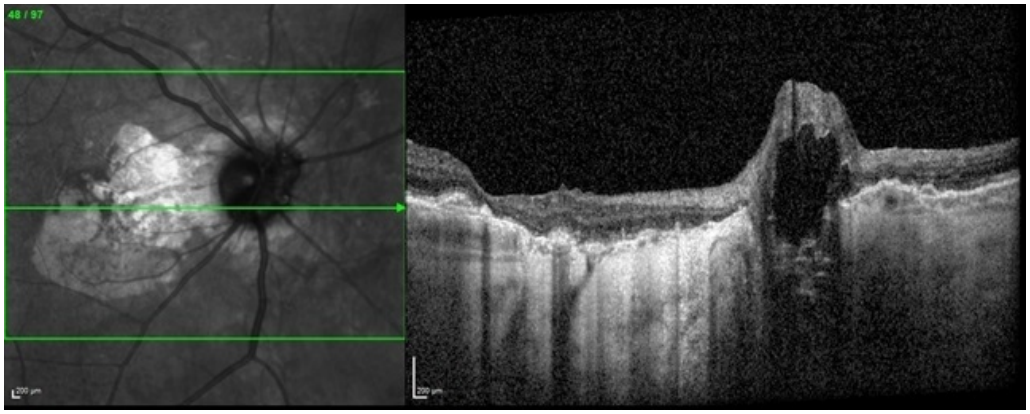

Figura 2.3: Ejemplo de imagen SLO (izquierda) y B-scan (derecha)

capas de la retina. En resumen, un B-Scan se corresponde con una línea horizontal o vertical de una imagen SLO (en la figura 2.3, a la izquierda, se puede observar una imagen SLO con una línea verde que indica el B-scan presentado en la imagen de la derecha) y un A-Scan con un punto de dicha línea.

Combinando la información de todos los B-scan que genera una tomografía de coherencia óptica es posible construir un modelo de tres dimensiones de las capas de la retina.

## 2.2. Momentos geométricos invariantes

Gracias a la información almacenada en una tomografía de coherencia óptica, es posible generar un modelo tridimensional que describa la capa de la retina, concretamente, la zona de la excavación de la papila (también llamada cúpula) que, como se ha explicado en la sección 2.1, permite diagnosticar el glaucoma. Mediante el estudio de esta depresión es posible conocer si el paciente se encuentra sano o padece la enfermedad (ya que esta neuropatía provoca el aumento de esta concavidad situada en el punto ciego de la retina). Sin embargo, dado que en la excavación de la papila se pueden observar distintas capas con una OCT, se elige aquella que se encuentra a modo de frontera entre la retina y el humor vítreo llamada membrana limitante interna (cuyas siglas son *ILM* debido a su nomenclatura en inglés *Inner Limiting Membrane*). De esta manera, se observa una superficie que se corresponde con la ILM y la excavación de la papila en ella. Cabe destacar que la capa de la ILM, no solo recubre la cúpula, sino que se encuentra sobre la retina. Por ello, y como la OCT nos ofrece "una imagen" tanto de la concavidad de la papila óptica como de la retina a su alrededor, basta con estudiar dicha membrana que se encontrará presente en todas las zonas capturadas por la tomografía.

Para describir dicha superficie se está interesado en encontrar un conjunto de momentos geométricos que sean invariantes ante semejanzas, que son transformaciones composición de rotaciones, traslaciones y reflexiones. La razón por la que se requiere esta propiedad de invarianza, es que, como se ha explicado, guiándose por la imagen SLO, el médico que lleva a cabo la OCT ajusta manualmente la máquina para capturar la papila óptica, por lo que esto puede ocasionar que la excavación no se encuentre siempre en el mismo lugar del espacio tridimensional.

Este tipo de momentos geométricos invariantes fueron presentados en dos dimensiones por primera vez por Hu (1962) para ser aplicados al reconocimiento de caracteres. Sin embargo, su artículo fue el precursor de muchos otros, como el redactado por los investigadores

Reddi (1981) quienes aprovecharon los momentos radial y angular para la identificación de imágenes.

Más adelante, la investigación de Sadjadi y Hall (1980) extendió la definición de estas entidades matemáticas a las superficies tridimensionales, tema que, desde entonces, ha suscitado mucho interés en la comunidad científica. De hecho, con el objetivo de construir momentos geométricos invariantes de tres dimensiones se han utilizado diversos métodos como un conjunto de resultados de la teoría de grupos que permiten generar momentos con coeficientes complejos (este es el caso de la investigación de los autores Lo y Don (1989)). También se han podido definir distintos momentos tridimensionales aplicando el cálculo tensorial (como Cyganski y Orr (1985) presentan en su artículo) y haciendo uso del teorema de Gauss para convertir un volumen integral en uno de superficie y disminuir con ello la complejidad computacional del momento (esta idea viene de la mano del investigador Li (1993)). Asimismo, se encuentran autores que, al igual que este estudio, muestran interés en aplicar ciertos momentos invariantes tridimensionales (a diferencia de lo que se presenta en este trabajo, ellos utilizan cálculo tensorial) a la investigación de imágenes médicas (Faber y Stokely, 1986).

En este trabajo, se hará uso de los momentos presentados por Xu y Li (2006), los cuales cumplen la propiedad de invarianza ante semejantes expresada anteriormente. Para explicar la definición dada por estos autores, es necesario mostrar algunos conceptos previos que facilitarán tanto su comprensión como la presentación de su formalización. Estas herramientas matemáticas se detallan a continuación.

Supóngase que $\varphi(u,v) = (x(u,v), y(u,v), z(u,v))$ es una superficie paramétrica en $\mathbb{R}^3$ tal que $\varphi : D \to S \subset \mathbb{R}^3$ con $D$ el dominio de definición de $(u,v)$ un subconjunto de $\mathbb{R}^2$. Entonces se definen los momentos de orden $k+m+n$ a través de las superficies integrales definidas sobre el área de S, es decir:

$$\begin{aligned} M_{kmn} &= \int_S x^k y^m z^n \varrho(x,y,z) ds \\ &= \int_D x^k(u_0,v_0) y^m(u_0,v_0) z^n(u_0,v_0) \varrho(x(u_0,v_0), y(u_0,v_0), z(u_0,v_0)) \sqrt{EG-F^2} du_0 dv_0 \end{aligned}$$

donde $E = x_u^2 + y_u^2 + z_u^2$, $G = x_v^2 + y_v^2 + z_v^2$ y $F = x_u x_v + y_u y_v + z_u z_v$ (aquí se utiliza la siguiente notación para denotar la derivada parcial: $x_u = (\partial x/\partial u)(u_0, v_0)$ con $(u_0, v_0) \in D$) son los coeficientes de la primera forma fundamental de $S$ (Rodríguez Sanjurjo y Ruiz Sancho, 2018, capítulo 5, páginas 57-59) y $\varrho(x,y,z)$ la función de densidad definida en la superficie. De hecho, la razón por la que utilizamos las variables $(u_0, v_0)$ en la última integral, es para diferenciarlas claramente de la dirección sobre la que se aplica la derivada parcial.

Con el concepto de momento de orden $k+m+n$, resulta bastante sencillo el cálculo del centroide de la superficie tridimensional haciendo uso de los momentos de orden cero y uno con las coordenadas que se muestran a continuación:

$$\overline{x} = \frac{M_{100}}{M_{000}}, \overline{y} = \frac{M_{010}}{M_{000}}, \overline{z} = \frac{M_{001}}{M_{000}}$$

Con el centroide, se puede introducir el concepto de momentos centrales de orden $k+m+n$:

$$\overline{M}_{kmn} = \int_S (x-\overline{x})^k (y-\overline{y})^m (z-\overline{z})^n \varrho(x,y,z) ds$$

Como afirman Xu y Li (2006), estos momentos centrales son invariantes ante traslaciones. De hecho, estos investigadores también presentan un momento invariante a reflexiones

y demuestran esta propiedad. No obstante, como se menciona anteriormente, es conveniente poseer momentos invariantes ante semejanzas (lo que significa que es necesario que también sean invariantes ante las rotaciones), ya que estos nos permitirán estudiar la superficie descrita por la ILM con independencia de la resolución a la que se obtenga el modelo tridimensional, la posición en el espacio de la excavación y otros factores que podrían ser determinantes e influyentes sobre los resultados del estudio en caso de no poseer esta propiedad de invarianza. A continuación, para concluir esta sección, se presentan siete momentos geométricos definidos por Xu y Li (2006) y que sí poseen el requerimiento de ser invariantes ante dicho conjunto de transformaciones:

$$M_0 := \overline{M}_{000}$$

$$M_1 := \frac{1}{\overline{M}_{000}^3}(\overline{M}_{400} + \overline{M}_{040} + \overline{M}_{004} + 2\overline{M}_{220} + 2\overline{M}_{202} + 2\overline{M}_{022})$$

$$\begin{aligned} M_2 := \frac{1}{\overline{M}_{000}^6}(&\overline{M}_{400}\overline{M}_{040} + \overline{M}_{400}\overline{M}_{004} + \overline{M}_{004}\overline{M}_{040} + 3\overline{M}_{220}^2 + 3\overline{M}_{202}^2 + 3\overline{M}_{022}^2 \\ &- 4\overline{M}_{103}\overline{M}_{301} - 4\overline{M}_{130}\overline{M}_{310} - 4\overline{M}_{013}\overline{M}_{031} + 2\overline{M}_{022}\overline{M}_{202} + 2\overline{M}_{022}\overline{M}_{220} \\ &+ 2\overline{M}_{220}\overline{M}_{202} + 2\overline{M}_{022}\overline{M}_{400} + 2\overline{M}_{004}\overline{M}_{220} + 2\overline{M}_{040}\overline{M}_{202} - 4\overline{M}_{103}\overline{M}_{121} \\ &- 4\overline{M}_{130}\overline{M}_{112} - 4\overline{M}_{013}\overline{M}_{211} - 4\overline{M}_{121}\overline{M}_{301} - 4\overline{M}_{310}\overline{M}_{112} - 4\overline{M}_{013}\overline{M}_{211} \\ &+ 4\overline{M}_{211}^2 + 4\overline{M}_{112}^2 + 4\overline{M}_{121}^2) \end{aligned}$$

$$\begin{aligned} M_3 := \frac{1}{\overline{M}_{000}^6}(&\overline{M}_{400}^2 + \overline{M}_{040}^2 + \overline{M}_{004}^2 + 4\overline{M}_{130}^2 + 4\overline{M}_{103}^2 + 4\overline{M}_{013}^2 + 4\overline{M}_{031}^2 + 4\overline{M}_{310}^2 + 4\overline{M}_{301}^2 \\ &+ 6\overline{M}_{220}^2 + 6\overline{M}_{202}^2 + 6\overline{M}_{022}^2 + 12\overline{M}_{112}^2 + 12\overline{M}_{121}^2 + 12\overline{M}_{211}^2) \end{aligned}$$

$$\begin{aligned} M_4 := \frac{1}{\overline{M}_{000}^5}(&\overline{M}_{300}^2 + \overline{M}_{030}^2 + \overline{M}_{003}^2 + 3\overline{M}_{120}^2 + 3\overline{M}_{102}^2 + 3\overline{M}_{012}^2 + 3\overline{M}_{021}^2 + 3\overline{M}_{210}^2 + 3\overline{M}_{201}^2 \\ &+ 6\overline{M}_{111}^2) \end{aligned}$$

$$\begin{aligned} M_5 := \frac{1}{\overline{M}_{000}^5}(&\overline{M}_{300}^2 + \overline{M}_{030}^2 + \overline{M}_{003}^2 + \overline{M}_{120}^2 + \overline{M}_{102}^2 + \overline{M}_{012}^2 + \overline{M}_{021}^2 + \overline{M}_{210}^2 + \overline{M}_{201}^2 \\ &+ 2\overline{M}_{300}\overline{M}_{120} + 2\overline{M}_{300}\overline{M}_{102} + 2\overline{M}_{102}\overline{M}_{120} + 2\overline{M}_{003}\overline{M}_{201} + 2\overline{M}_{003}\overline{M}_{021} \\ &+ 2\overline{M}_{021}\overline{M}_{201} + 2\overline{M}_{030}\overline{M}_{012} + 2\overline{M}_{030}\overline{M}_{210} + 2\overline{M}_{012}\overline{M}_{210}) \end{aligned}$$

$$\begin{aligned} M_6 := \frac{1}{\overline{M}_{000}^5}[&\overline{M}_{200}(\overline{M}_{400} + \overline{M}_{220} + \overline{M}_{202}) + \overline{M}_{020}(\overline{M}_{220} + \overline{M}_{040} + \overline{M}_{022}) \\ &+ \overline{M}_{002}(\overline{M}_{202} + \overline{M}_{022} + \overline{M}_{004}) + 2\overline{M}_{110}(\overline{M}_{310} + \overline{M}_{130} + \overline{M}_{112}) \\ &+ 2\overline{M}_{101}(\overline{M}_{301} + \overline{M}_{121} + \overline{M}_{103}) + 2\overline{M}_{011}(\overline{M}_{211} + \overline{M}_{031} + \overline{M}_{013})] \end{aligned}$$

Estos siete momentos geométricos que se han expresando formalmente, son utilizados, en este estudio, para describir la superficie que genera la ILM aprovechando los resultados obtenidos por Rodríguez (1/2017 - 12/2019). En concreto, para estudiar la excavación que en esta membrana se observa. Por esta razón, se aplican estos momentos tanto a la ILM completa como la zona de la membrana en la que se encuentra la excavación (obteniendo momentos resultantes distintos en cada caso, ya que la cúpula representa una fracción de toda la ILM suficiente pequeña como para que estos varíen). Dado que la tarea de recortar la imagen en tres dimensiones de la ILM, para quedarse exclusivamente con la cúpula, es una labor que se debe comprobar su corrección manualmente (dada la gigantesca casuística de los B-scans y el amplio número de errores en ellos), no ha sido posible, por limitación de recursos temporales, extraer la depresión de todas las superficies generadas a partir de

las tomografías facilitadas por los expertos en oftalmología. Como consecuencia de ello, se cuenta con un conjunto menor de tuplas de los siete momentos restringidos a la excavación.

Con la posibilidad en mente de que en un futuro se consiga una muestra mayor de imágenes tomográficas, previo al cálculo de estos momentos, se ha escalado el modelo tridimensional, de manera que se obtenga dicha imagen en tres dimensiones a tamaño real independientemente del número de B-scans y A-scans utilizados. Así, si las nuevas muestras obtenidas de la OCT tuvieran una configuración distinta, podrían ser añadidas (escalando la imagen previamente) al conjunto de datos. De esta forma, se cuenta tanto con los momentos de la imagen original (los momentos *raw* o en bruto) como con los de la imagen escalada. En el análisis de los datos, siempre se hace uso de estos últimos para poder extrapolar y extender los resultados y conclusiones obtenidas incluso a tomografías con configuraciones distintas.

Capítulo 3

# Presentación del conjunto de datos

*"Los datos por sí solos no tienen valor, son sólo masas de números o palabras."*
— Total Value Optimization: Transforming Your Global Supply Chain Into a Competitive Weapon
Steven J. Bowen (2017)

Antes de proceder con el análisis de los datos haciendo uso de técnicas de aprendizaje automático, resulta imprescindible comprender las distintas variables con las que cuenta el sistema y las principales características que presentan. Por esta razón, en este capítulo se lleva a cabo una exhaustiva presentación de las variables que describen el conjunto de datos (véase la sección 3.1) que dan lugar a conocer tanto la configuración de la máquina de OCT con la que se realizó la prueba como detalles del paciente.

A continuación se extraen las características más superficiales de los momentos geométricos invariantes mediante el uso de técnicas estadísticas básicas. De esta forma, se estudia la consistencia de estos invariantes cuando se lleva a cabo la misma prueba el mismo día (aunque en horas distintas) al mismo paciente (véase la sección 3.2), se calcula los estadísticos descriptivos más comunes (como la media y la desviación típica) de los distintos momentos (véase la sección 3.3) y se evalúan las relaciones de correlación y dispersión entre cada par de invariantes (véase la sección 3.4). Estos datos por sí solos, no reflejarán brillantes conclusiones, pero son de gran utilidad para más tarde aplicar las técnicas de aprendizaje automático.

## 3.1. Variables del sistema

Dentro del conjunto de datos se pueden observar un amplio número de variables que definen cada imagen tomográfica. Concretamente, se cuenta con 53 variables distintas, de las cuales 28 se corresponden con los momentos geométricos. Este amplio número de dimensiones que se corresponden con los invariantes explicados en la sección 2.2, se debe a que se han calculado tanto de la imagen escalada a tamaño real como de la imagen original. Además, de cada una de ellas se han evaluado los momentos geométricos sobre toda la ILM y restringiéndolos a la excavación presente en la papila óptica. Por lo tanto, se han calculado cuatro veces el conjunto de siete invariantes explicado anteriormente, con lo que se obtienen las 28 características mencionadas. No obstante, dado que los momentos geométricos calculados sobre la imagen escalada permiten aprovechar y extender los resultados de este estudio en el caso de obtener imágenes de OCT con configuraciones distintas

(como por ejemplo un número diferente de B-Scans), a lo largo de todo el análisis se hace uso de estos últimos en lugar de los que se han evaluado a partir de la imagen original.

En lo que se refiere a las 25 variables que no se corresponden con los momentos geométricos, la mayor parte de ellos describen la configuración con la que se llevó a cabo la OCT (como la versión del formato del archivo exportado por la máquina de OCT, el tipo de patrón que ha utilizado la exploración o el número de B-Scans, entre otras aclaraciones técnicas). Otro considerable conjunto pormenorizan detalles del paciente como su fecha de nacimiento, el ojo sobre el que está hecha la prueba o su identificador. El único campo incluido en este estudio es aquel que, utilizando un valor booleano, indica si el paciente pertenece al grupo de control o padece la enfermedad. Toda la información sobre estas variables del sistema se encuentra condensada y especificada en la tabla 3.1.

Sin entrar en detalles técnicos de las distintas variables que no sean relevantes para este estudio, cabe destacar que inicialmente se implementó, utilizando el lenguaje Python y beneficiándose de las herramientas facilitadas por las librerías *pandas* y *numpy*, una clase llamada *MyDataset* que almacena un *DataFrame* (McKinney et al., 2011) y extrae como metadatos del *dataset* aquellas variables que se mantengan siempre constantes (se corresponden con todas aquellas que poseen un valor definido en la última columna de la tabla 3.1).

Aunque, como se ha mencionado, en este trabajo solo se presta atención al comportamiento de los distintos momentos geométricos, el conocimiento de estas variables hace posible formarse una idea del tipo de datos con los que se están trabajando.

Por último, es importante mencionar que en el conjunto de datos se pueden detectar algunas tomografías realizadas al mismo paciente en dos ocasiones el mismo día. Dado que el estado de la excavación y de la ILM no varía en un periodo tan corto de tiempo (a menos que el paciente haya sido sometido a algún tipo de medicación que altere la cúpula), este hecho hace posible la comprobación de consistencia de los momentos geométricos definidos como se presenta a continuación.

## 3.2. Consistencia de los momentos geométricos

Aprovechando la existencia de seis pares de tomografías de coherencia óptica realizadas al mismo paciente en un mismo día, una de las preguntas que pueden abordarse es si los momentos geométricos serán consistentes y denotarán una clara diferencia entre la pareja de la misma persona y el resto de OCT. Para ser capaces de cuantificar la similitud entre los momentos geométricos de las dos pruebas del mismo paciente y la diferencia que posee con respecto al resto de personas, se decide utilizar tanto la distancia euclídea como el error cuadrático medio. Con el fin de poder evaluar ambas métricas, se tomarán vectores de siete elementos que se corresponderán con los siete invariantes aplicados a la imagen escalada. Asimismo, se estudia la consistencia dos veces: tanto en el caso en que los momentos describen toda la ILM, como aquel en que se restringen a la excavación presente en la papila óptica.

Como se ha mencionado, se utilizará tanto la distancia euclídea entre dos OCT (que cada uno viene determinado por los siete momentos) como el error cuadrático medio, el cual viene dado por:

$$ECM(a,b) := \frac{1}{7}\sum_{i=0}^{6}(M_i^a - M_i^b)^2$$

donde $M_i^a$ se corresponde con el i-ésimo momento geométrico de la OCT a.

| **Nombre** | **Descripción** | **Tipo de dato** | **Valor constante** |
|---|---|---|---|
| version | Número de versión del formato de archivo exportado por la máquina de la OCT. | str | HSF-OCT-103 |
| scanPattern | Indica el tipo de patrón de la exploración:<br>0. Patrón desconocido.<br>1. Exploración de línea única (un solo B-Scan).<br>2. Exploración circular (un solo B-Scan).<br>3. Exploración de volumen en modo ART.<br>4. Exploración rápida de volumen.<br>5. Exploración radial (o patrón de estrella). | int | 3 |
| sizeXSlo | Anchura (en píxeles) de la imagen SLO. | int | 768 |
| sizeYSlo | Altura (en píxeles) de la imagen SLO. | int | 768 |
| fieldSizeSlo | Tamaño del área horizontal de la imagen SLO en grados. | int | 30 |
| numBScan | Número de B-scans en la exploración OCT. | int | 97 |
| sizeX | Número de A-scans en cada B-scan, es decir, el ancho de cada B-scan en píxeles. | int | 384 |
| sizeZ | Número de muestras en un A-scan, es decir, el alto de cada B-scan en píxeles. | int | 496 |
| scaleZ | Altura de un píxel de un B-scan medida en mm. | float | 0.0038717 |
| numSeg | Número de vectores de segmentación. | int | 3 |
| patientID | Identificador de paciente. | str | - |
| pid | Identificador interno del paciente. | int | - |
| Patient | Nombre del archivo exportado por la máquina de OCT que posee la siguiente estructura <Primer apellido del paciente>_<Segundo apellido>_<Inicial del nombre>_ <número único identificador de la OCT>.vol | str | - |
| dob | Fecha de nacimiento del paciente. | str | - |
| scanPosition | Ojo examinad: "OS" para el ojo izquierdo y "OD" para el ojo derecho. | str | - |
| examTime | Momento en el que se llevó a cabo la OCT. | str | - |
| id | Identificador único de la OCT. Coincide con el último número del campo Patient | int | - |
| scanFocus | Foco del escáner. | float | - |
| scaleXSlo | Anchura de un píxel de la imagen SLO en mm. | float | - |
| scaleYSlo | Altura de un píxel de la imagen SLO en mm. | float | - |
| distance | Distancia entre dos B-scans adyacentes en mm. | float | - |
| scaleX | Anchura del píxel de un B-scan medida en mm. | float | - |
| maximum_depth | Profundidad máxima de las segmentaciones. | float | - |
| bscan_qs | Calidad de cada uno de los B-scans realizados. | list | - |
| glaucoma | Indicador de si el paciente padece la enfermedad. | bool | - |

Tabla 3.1: Variables del sistema no relacionadas con los momentos geométricos

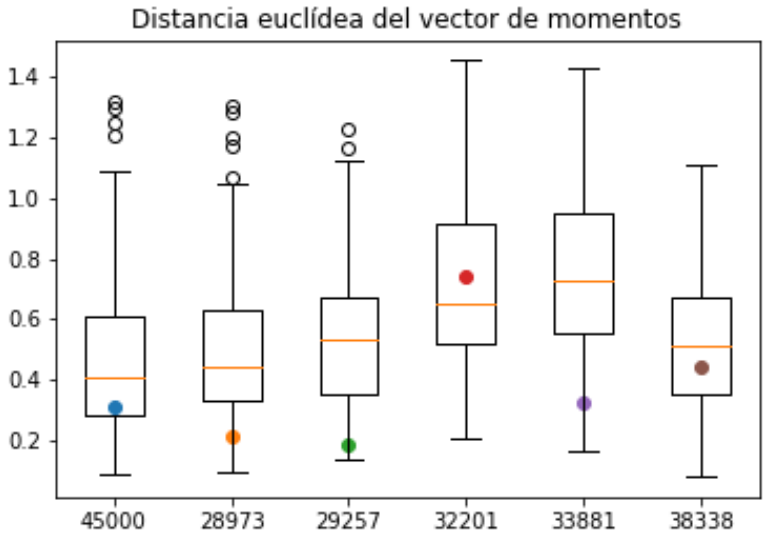


Figura 3.1: Distribución de distancia euclídea de los momentos sobre toda la ILM.

Así, mediante el uso de las librerías *numpy*, *pandas* y *matplotlib* de Python, se implementa el cálculo de las dos distancias desde un vector de momentos de una OCT, de un paciente que posee dos en un mismo día, hasta el resto de OCT con pacientes distintos. Previamente, se escalan los diferentes invariantes para evitar que la posible existencia de grandes diferencias de magnitud entre los valores obtenidos de los momentos (como se demuestra que ocurre en la sección 3.3), influencie excesivamente en las dos métricas utilizadas.

A continuación se estudia la distribución de dicho conjunto de distancias y se representa en un gráfico de cajas (Potter et al., 2006) como el que se muestra en la figura 3.1. También se calculan los estadísticos descriptivos más comunes (media, desviación típica, cuartiles, valor mínimo y valor máximo) de dicho conjunto de distancias.

Comenzando por el caso en que los momentos geométricos describen toda la ILM, se puede observar la distribución descrita por la figura 3.1. Dicha imagen representa cómo se distribuyen la distancia euclídea desde el paciente con *pid* que aparece en el eje de abscisas hasta el resto de personas (aparecen seis diagramas de cajas ya que se cuenta con seis OCT duplicadas de un mismo paciente el mismo día, pero en horas distintas). Dicho diagrama indica la mediana de distancias con el segmento horizontal anaranjado dentro de los distintos rectángulos, cuyos extremos superior e inferior se corresponde con el cuartil superior e inferior, respectivamente. Las líneas que se extienden paralelas a las cajas se conocen como "bigotes". Estos segmentos permiten identificar las observaciones atípicas, que son aquellas que quedan fuera del espacio comprendido entre los límites inferior y superior de los bigotes y en la figura 3.1 vienen representadas a través de puntos huecos. Por último se pueden observar puntos de colores en cada uno de los seis diagramas. Estos elementos representan la distancia a la que se encuentran los vectores de momentos duplicados del mismo paciente. Como puede verse, en la mayoría de ocasiones, los momentos no son consistentes, ya que es posible observar vectores de otros pacientes que se encuentran a menor distancia que los del mismo paciente. Incluso hay un caso (como la persona con pid 32201) cuya distancia a su mismo momento se encuentra por encima de la mediana. Con el fin de asegurarse de que este comportamiento descrito en el gráfico refleja correctamente la

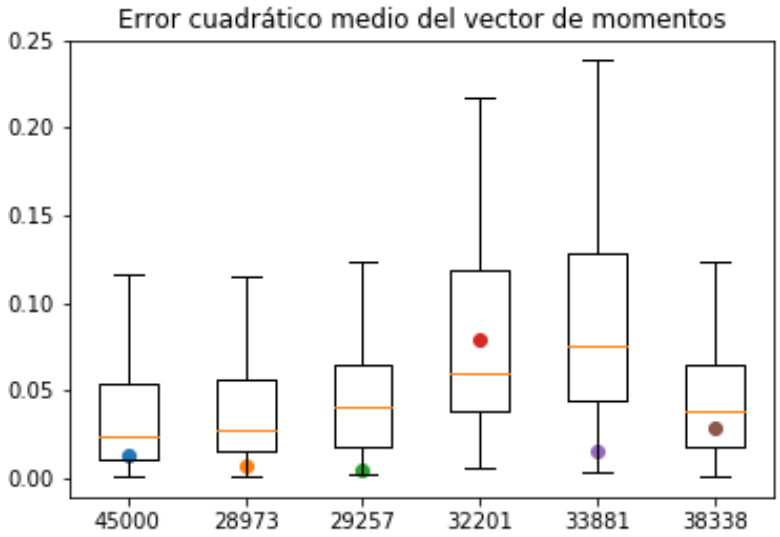


Figura 3.2: Distribución del ECM de los momentos sobre toda la ILM.

relación entre las distancias, se calculan los estadísticos descriptivos más comunes.

En efecto, si se observan los valores que presenta la tabla 3.2, se confirma lo deducido a partir de la figura 3.1: los momentos geométricos no presentan consistencia con la distancia euclídea cuando describen toda la ILM. Dadas las similitudes entre las expresiones de la distancia euclídea y el error cuadrático medio, cabe esperar que tampoco se presente la propiedad de consistencia cuando se lleva a cabo el mismo análisis con este último. Sin mostrar la tabla de estadísticos descriptivos más comunes, a continuación, se presenta el diagrama de cajas y bigotes utilizando el error cuadrático medio (véase la figura 3.2).

En este caso, dado que dificultaban la visualización de las cajas por su gran dispersión, en la figura 3.2 no se muestran los puntos atípicos de la distribución del ECM. Como puede observarse comparando dicha imagen con la figura 3.1, se presenta un comportamiento muy parecido respecto a la distancia entre los vectores de un mismo paciente en relación con los de otros pacientes. Por lo tanto, como obtenemos las mismas conclusiones, podemos negar la existencia de consistencia en los momentos geométricos invariantes a la hora de describir toda la ILM.

A pesar de esta conclusión, queda por estudiar si se presenta la propiedad de consisten-

| | **45000** | **28973** | **29257** | **32201** | **33881** | **38338** |
|---|---|---|---|---|---|---|
| **media** | 0.490 | 0.511 | 0.547 | 0.722 | 0.729 | 0.531 |
| **desviación típica** | 0.285 | 0.269 | 0.252 | 0.256 | 0.297 | 0.257 |
| **valor mínimo** | 0.087 | 0.095 | 0.137 | 0.203 | 0.166 | 0.080 |
| **25 %** | 0.280 | 0.333 | 0.353 | 0.518 | 0.555 | 0.355 |
| **50 %** | 0.409 | 0.440 | 0.532 | 0.649 | 0.728 | 0.515 |
| **75 %** | 0.612 | 0.627 | 0.675 | 0.912 | 0.949 | 0.674 |
| **valor máximo** | 1.318 | 1.305 | 1.226 | 1.454 | 1.430 | 1.109 |
| **distancia al duplicado** | 0.307 | 0.214 | 0.185 | 0.744 | 0.323 | 0.445 |

Tabla 3.2: Estadísticos descriptivos más comunes de la distancia euclídea entre los momentos sobre toda la ILM

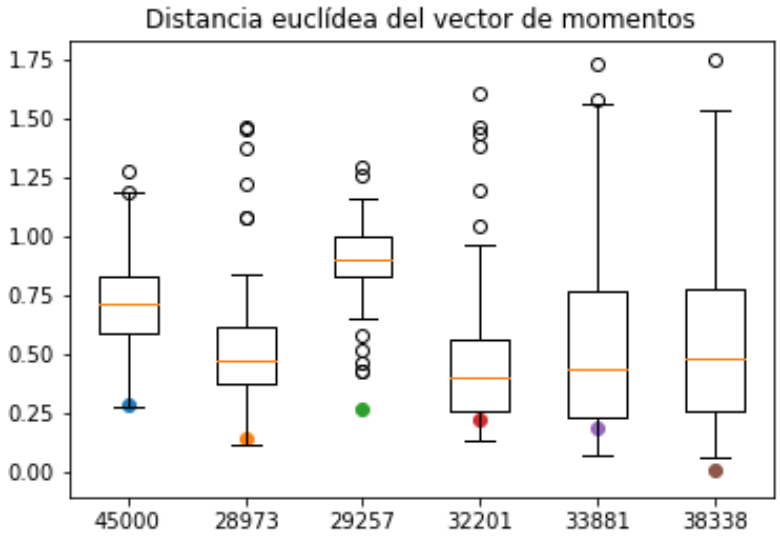


Figura 3.3: Distribución de distancia euclídea de los momentos restringidos a la excavación.

cia cuando los momentos geométricos se restringen a la excavación de la ILM. Para ello, se hace uso, al igual que en el caso anterior, tanto de la distancia euclídea como del error cuadrático medio. En el caso de la primera, nos encontramos con la distribución presente en la figura 3.3. En ella se observan claras mejorías en cuanto a que disminuye la distancia de los momentos de un mismo paciente con relación al del resto de personas (incluso se presentan dos pacientes, con pid 29257 y 38338, cuya distancia al duplicado es menor que el resto de elementos). Hablamos de resultados mejores ya que todas las muestras duplicadas se encuentran por debajo del primer cuartil. Sin embargo, también puede verse que en algunos casos continúan existiendo vectores de momentos de otros pacientes "más similares" que los del propio. Este hecho, combinado con que el tamaño del conjunto de datos de los momentos geométricos restringidos a la cúpula es menor, da lugar a la sospecha de que la no se presenta la consistencia y que los resultados son análogos a los del caso anterior.

Para confirmar estas deducciones, resulta conveniente detallar los estadísticos descriptivos más comunes también de este conjunto. Estos se presentan en la tabla 3.3. Como puede observarse, a excepción del tercer y último paciente, en todos los demás casos encontramos una OCT de otra persona que se asemeja más que la del mismo individuo. Estos

| | **45000** | **28973** | **29257** | **32201** | **33881** | **38338** |
|---|---|---|---|---|---|---|
| **media** | 0.718 | 0.561 | 0.896 | 0.503 | 0.555 | 0.575 |
| **desviación típica** | 0.211 | 0.300 | 0.186 | 0.353 | 0.417 | 0.405 |
| **valor mínimo** | 0.284 | 0.123 | 0.433 | 0.140 | 0.074 | 0.070 |
| **25 %** | 0.591 | 0.380 | 0.833 | 0.262 | 0.236 | 0.261 |
| **50 %** | 0.717 | 0.479 | 0.902 | 0.405 | 0.436 | 0.481 |
| **75 %** | 0.832 | 0.616 | 0.997 | 0.562 | 0.772 | 0.782 |
| **valor máximo** | 1.273 | 1.464 | 1.298 | 1.606 | 1.729 | 1.745 |
| **distancia al duplicado** | 0.288 | 0.147 | 0.269 | 0.223 | 0.194 | 0.010 |

Tabla 3.3: Estadísticos descriptivos más comunes de la distancia euclídea entre los momentos restringidos a la cúpula

datos nos confirman la inexistencia de consistencia también en el caso en que los momentos geométricos invariantes se restringen a la zona de la ILM donde se encuentra la excavación.

Aunque se ha comprobado, no es necesario detallar la distribución obtenida haciendo uso del error cuadrático medio, ya que la semejanza de su expresión con la de la distancia euclídea produce resultados similares (como se vio en el caso en que los momentos eran aplicados a toda la ILM). De esta forma, y con los datos presentados, es posible contestar a la pregunta que se formulaba al inicio de esta sección sobre la consistencia de los invariantes. Con la muestra con la que se cuenta (y suponiendo que la cúpula no ha sido sometida a ningún tratamiento médico, lo cual se desconoce), los momentos geométricos no presentan consistencia cuando se aplican a dos OCT distintas del mismo individuo. No obstante, este hecho podría no ejercer influencia sobre otras cuestiones como el diagnóstico automático de la enfermedad, ya que podría ocurrir que los pacientes más parecidos que los del propio individuo se encuentre en el mismo conjunto (grupo de control o paciente de glaucoma).

## 3.3. Estadísticos descriptivos de los momentos geométricos

Una buena práctica antes de comenzar cualquier análisis de datos es obtener los estadísticos descriptivos más comunes del conjunto de características que describen el sistema. De esta forma, podrían salir a relucir detalles que den lugar a sospechar alguna propiedad presente en el conjunto de datos. Asimismo, es posible justificar ciertos comportamientos en las técnicas de aprendizaje automático a través de ellos. Probablemente, las conclusiones obtenidas a partir de los estadísticos descriptivos no sean determinantes, pero pueden ser de gran utilidad posteriormente. Por esta razón, se analizan los valores obtenidos y se desarrollan conclusiones sobre ellos a lo largo de esta sección.

Con el objetivo de calcular estos estadísticos descriptivos, se hace uso de las funciones implementadas en los *DataFrame* de la librería de Python *pandas*. De manera que basta con poseer el conjunto de datos en dicho formato para generar estos valores fácilmente.

En primer lugar, se tratan los estadísticos descriptivos más comunes de los momentos geométricos, sobre la imagen escalada, restringidos a la excavación de la ILM. Se debe tener en cuenta que este conjunto de datos es menor que si no se aplica dicha restricción (concretamente posee 56 vectores compuestos por los siete momentos geométricos). Los estadísticos dados por este sistema pueden observarse en la tabla 3.4.

Resulta interesante destacar que la desviación típica de todos los momentos se distingue por ser bastante representativa en lo que se refiere al orden de magnitud de cada uno de ellos. Este hecho puede levantar la sospecha de que todos los invariantes son muy variables y de que de ello depende el diagnóstico de la enfermedad. Para tratar de dirimir esta cuestión se deben estudiar los estadísticos descriptivos por separado tanto del grupo de

| | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ | $M_6$ |
|---|---|---|---|---|---|---|---|
| **media** | 4.19 | 0.032 | 0.012 | $5 \cdot 10^{-4}$ | $7 \cdot 10^{-5}$ | $3 \cdot 10^{-5}$ | 0.003 |
| **desviación típica** | 0.969 | 0.009 | 0.011 | $5 \cdot 10^{-4}$ | $6{,}4 \cdot 10^{-5}$ | $2 \cdot 10^{-5}$ | 0.002 |
| **valor mínimo** | 2.648 | 0.018 | 0.002 | $10^{-4}$ | $5 \cdot 10^{-6}$ | $7 \cdot 10^{-7}$ | $9 \cdot 10^{-4}$ |
| **25 %** | 3.567 | 0.026 | 0.006 | $3 \cdot 10^{-4}$ | $4 \cdot 10^{-5}$ | $2 \cdot 10^{-5}$ | 0.002 |
| **50 %** | 3.903 | 0.029 | 0.008 | $4 \cdot 10^{-4}$ | $5 \cdot 10^{-5}$ | $3 \cdot 10^{-5}$ | 0.002 |
| **75 %** | 4.864 | 0.036 | 0.016 | $6 \cdot 10^{-4}$ | $8 \cdot 10^{-5}$ | $4 \cdot 10^{-5}$ | 0.003 |
| **valor máximo** | 6.892 | 0.06 | 0.06 | 0.002 | $4 \cdot 10^{-4}$ | $10^{-4}$ | 0.008 |

Tabla 3.4: Estadísticos descriptivos de los momentos restringidos a la excavación

| | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ | $M_6$ |
|---|---|---|---|---|---|---|---|
| **media** | 4.046 | 0.035 | 0.012 | $7 \cdot 10^{-4}$ | $6 \cdot 10^{-5}$ | $2 \cdot 10^{-5}$ | 0.003 |
| **desviación típica** | 0.89 | 0.01 | 0.01 | $6 \cdot 10^{-4}$ | $4 \cdot 10^{-5}$ | $10^{-5}$ | 0.002 |
| **valor mínimo** | 2.648 | 0.023 | 0.003 | $2 \cdot 10^{-4}$ | $5 \cdot 10^{-6}$ | $7 \cdot 10^{-7}$ | 0.002 |
| **25 %** | 3.513 | 0.029 | 0.007 | $3 \cdot 10^{-4}$ | $2 \cdot 10^{-5}$ | $10^{-5}$ | 0.002 |
| **50 %** | 3.847 | 0.031 | 0.009 | $4 \cdot 10^{-4}$ | $5 \cdot 10^{-5}$ | $3 \cdot 10^{-5}$ | 0.002 |
| **75 %** | 4.64 | 0.038 | 0.011 | $7 \cdot 10^{-4}$ | $8 \cdot 10^{-5}$ | $3 \cdot 10^{-5}$ | 0.004 |
| **valor máximo** | 5.713 | 0.06 | 0.038 | 0.002 | $10^{-4}$ | $5 \cdot 10^{-5}$ | 0.008 |

Tabla 3.5: Estadísticos descriptivos del grupo de control en la excavación

control como de los pacientes de glaucoma. Sin embargo, basta observar la tabla 3.5, en la que se presentas los estadísticos de los momentos geométricos del grupo de control (el cual tiene un tamaño de 21 elementos), para desechar esta idea, ya que tanto la media como la desviación típica son muy similares en prácticamente todos los casos. Además también se parecen en gran medida el resto de los estadísticos descriptivos.

También es interesante señalar la gran diferencia en el orden de magnitud entre los distintos momentos que se hace patente en ambas tablas. Este hecho es importante debido a que obliga en muchas ocasiones (como cuando queremos ejecutar el algoritmo K-Medias como se presenta en la sección 4.1.3) a escalar los momentos, como se hizo cuando se presentó el estudio de consistencia de los invariantes. Por lo tanto es algo se tendrá en cuenta a lo largo de todo el trabajo.

Para ser capaces de afirmar con total certeza que la alta desviación típica presente en la tabla 3.4 no se debe a la diferencia entre personas que padecen la enfermedad y pacientes del grupo de control, se generan los estadísticos descriptivos del primer conjunto. Este grupo de datos posee un total de 35 muestras. Los resultados del cálculo se presentan en la tabla 3.6.

En efecto, como indicaban los datos de la tabla previa, ambos posee valores medios bastantes parecidos, si se tiene en cuenta en dicha comparación, la desviación típica de cada grupo de datos. Asimismo el resto de estadísticos resultan suficientemente similares como para que la distinción entre los conjuntos no resulte tan evidente. Una ventaja es que la gran diferencia en el orden de magnitud se mantiene y vuelve a producirse como era de esperar.

Por último, también es conveniente presentar los estadísticos descriptivos cuando los momentos geométricos describen toda la superficie de la ILM. Dado que este conjunto es significativamente mayor que el anterior (posee un total de 80 muestras), con los valores calculados en la tabla 3.7 las conclusiones son más relevantes que en el caso anterior.

| | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ | $M_6$ |
|---|---|---|---|---|---|---|---|
| **media** | 4.277 | 0.03 | 0.012 | $5 \cdot 10^{-4}$ | $8 \cdot 10^{-5}$ | $4 \cdot 10^{-5}$ | 0.002 |
| **desviación típica** | 1.016 | 0.009 | 0.011 | $5 \cdot 10^{-4}$ | $8 \cdot 10^{-5}$ | $3 \cdot 10^{-5}$ | 0.002 |
| **valor mínimo** | 2.842 | 0.018 | 0.002 | $10^{-4}$ | $2 \cdot 10^{-5}$ | $9 \cdot 10^{-6}$ | $9 \cdot 10^{-4}$ |
| **25 %** | 3.63 | 0.025 | 0.006 | $2 \cdot 10^{-4}$ | $4 \cdot 10^{-5}$ | $2 \cdot 10^{-5}$ | 0.002 |
| **50 %** | 3.945 | 0.028 | 0.008 | $3 \cdot 10^{-4}$ | $5 \cdot 10^{-5}$ | $3 \cdot 10^{-5}$ | 0.002 |
| **75 %** | 4.888 | 0.031 | 0.016 | $5 \cdot 10^{-4}$ | $7 \cdot 10^{-5}$ | $4 \cdot 10^{-5}$ | 0.003 |
| **valor máximo** | 6.892 | 0.057 | 0.06 | 0.002 | $4 \cdot 10^{-4}$ | $10^{-4}$ | 0.008 |

Tabla 3.6: Estadísticos descriptivos de los pacientes de glaucoma en la excavación

| | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ | $M_6$ |
|---|---|---|---|---|---|---|---|
| **media** | 18.463 | 0.038 | 0.98 | $7 \cdot 10^{-4}$ | $9 \cdot 10^{-5}$ | $7 \cdot 10^{-5}$ | 0.004 |
| **desviación típica** | 2.142 | 0.004 | 0.241 | $10^{-4}$ | $7 \cdot 10^{-5}$ | $6 \cdot 10^{-5}$ | 0.001 |
| **valor mínimo** | 14.855 | 0.029 | 0.49 | $4 \cdot 10^{-4}$ | $7 \cdot 10^{-7}$ | $6 \cdot 10^{-7}$ | 0.003 |
| **25 %** | 16.908 | 0.036 | 0.832 | $7 \cdot 10^{-4}$ | $4 \cdot 10^{-5}$ | $2 \cdot 10^{-5}$ | 0.003 |
| **50 %** | 18.407 | 0.038 | 0.964 | $8 \cdot 10^{-4}$ | $8 \cdot 10^{-5}$ | $5 \cdot 10^{-5}$ | 0.004 |
| **75 %** | 19.77 | 0.040 | 1.092 | $8 \cdot 10^{-4}$ | $10^{-4}$ | $10^{-4}$ | 0.004 |
| **valor máximo** | 26.059 | 0.047 | 2.071 | 0.001 | $3 \cdot 10^{-4}$ | $3 \cdot 10^{-4}$ | 0.005 |

Tabla 3.7: Estadísticos descriptivos de los momentos de toda la ILM

La primera apreciación que debe expresarse al mostrar los datos de la tabla 3.7, es que los valores de los momentos geométricos son significativamente mayores que en el caso en que se restringen a la excavación. Por el contrario, solo en tres de los siete invariantes se puede observar un trascendental aumento de la desviación típica. De esos tres casos (que son los momentos $M_0$, $M_2$ y $M_5$), en el primero no se trata de un incremento relevante en comparación con la media que se observa. En consecuencia a dicho aumento del valor del primer estadístico descriptivo, la mayoría de los valores mínimos se han incrementado de manera notable. Lo mismo se puede ver que ocurre con tanto los valores máximos como con gran parte de los cuartiles. También vuelve a denotarse una clara distinción de magnitud entre varios momentos, a la cual se tendrá que prestar atención en función de las circunstancias.

## 3.4. Correlación y dispersión de los momentos geométricos

Otra información que puede resultar de interés antes de comenzar con la aplicación de técnicas de aprendizaje automático, es el estudio de la relación (entre cada par) que presentan los distintos momentos geométricos. Para ello se calculará tanto el coeficiente de correlación de Pearson (Benesty et al., 2009) como se presentará la matriz de dispersión (en la que se relacionan gráficamente las distintas dimensiones representándolas por parejas).

### 3.4.1. Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson es una medida de la dependencia lineal entre dos variables aleatorias cuantitativas. Es posible calcular este coeficiente a partir de un estadístico muestral. Por lo tanto, si se posee una muestra de $N \in \mathbb{N}$ elementos del sistema (cada elemento viene determinado por una tupla de 7 coordenadas que se corresponden con cada momento geométrico), se contará precisamente con $N$ muestras de cada uno de los invariantes. Si se denota por $\{X_i^k\}_{i=1}^N$ el conjunto de muestras del momento geométrico k-ésimo (con $k \in \{0, \ldots, 6\}$), tenemos que el coeficiente de correlación de Pearson entre los invariantes $j$ y $k$ se define como:

$$r_{jk} = \frac{N \sum_{i=1}^N X_i^j X_i^k - \sum_{i=1}^N X_i^j \sum_{i=1}^N X_i^k}{\sqrt{N \sum_{i=1}^N (X_i^j)^2 - \left(\sum_{i=1}^N X_i^j\right)^2} \sqrt{N \sum_{i=1}^N (X_i^k)^2 - \left(\sum_{i=1}^N X_i^k\right)^2}}$$

A diferencia de la covarianza muestral, el coeficiente de correlación de Person es absolutamente independiente de la escala de medida de las variables. Por lo tanto, a pesar de haber observado en la sección 3.3 la gran variabilidad de órdenes de magnitud, no se requerirá escalar los momentos geométricos.

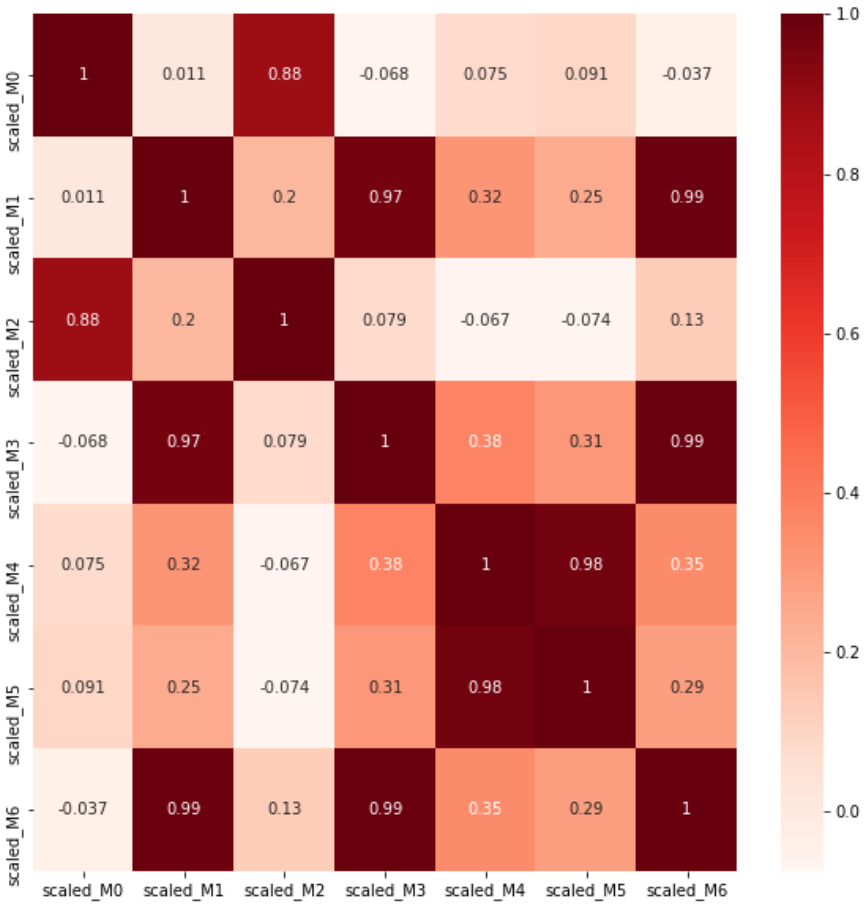


Figura 3.4: Coeficiente de correlación de Pearson de los momentos restringidos a la excavación

En un lenguaje menos formal, es posible definir el coeficiente de correlación de Pearson como un índice que puede ser utilizado para medir el grado de relación entre dos variables siempre y cuando estas sean cualitativas y continuas. Dicho coeficiente toma valores entre -1 y 1, donde 1 representa una correlación lineal positiva total, 0 indica la inexistencia de correlación lineal y -1 muestra una correlación lineal negativa total.

El resultado de calcular el coeficiente de correlación de Pearson sobre cada pareja de momentos geométricos cuando se restringen a la excavación se muestra en el mapa de calor representado en la figura 3.4. En él apenas se observan indicios de correlación lineal negativa, por lo que las celdas más claras indican inexistencia de correlación lineal. Además, se pueden deducir claras correlaciones a partir de dicha gráfica. Las más representativas son las que presenta el último momento geométrico $M_6$ con el segundo $M_1$ y el cuarto $M_3$, la evidente correlación entre el quinto invariante $M_4$ y el sexto $M_5$ y la linealidad observada entre el segundo momento $M_1$ y el cuarto $M_3$. Por otro lado, también destaca un indicio de correlación entre el primero invariante $M_0$ y el tercero $M_3$, por su elevado coeficiente

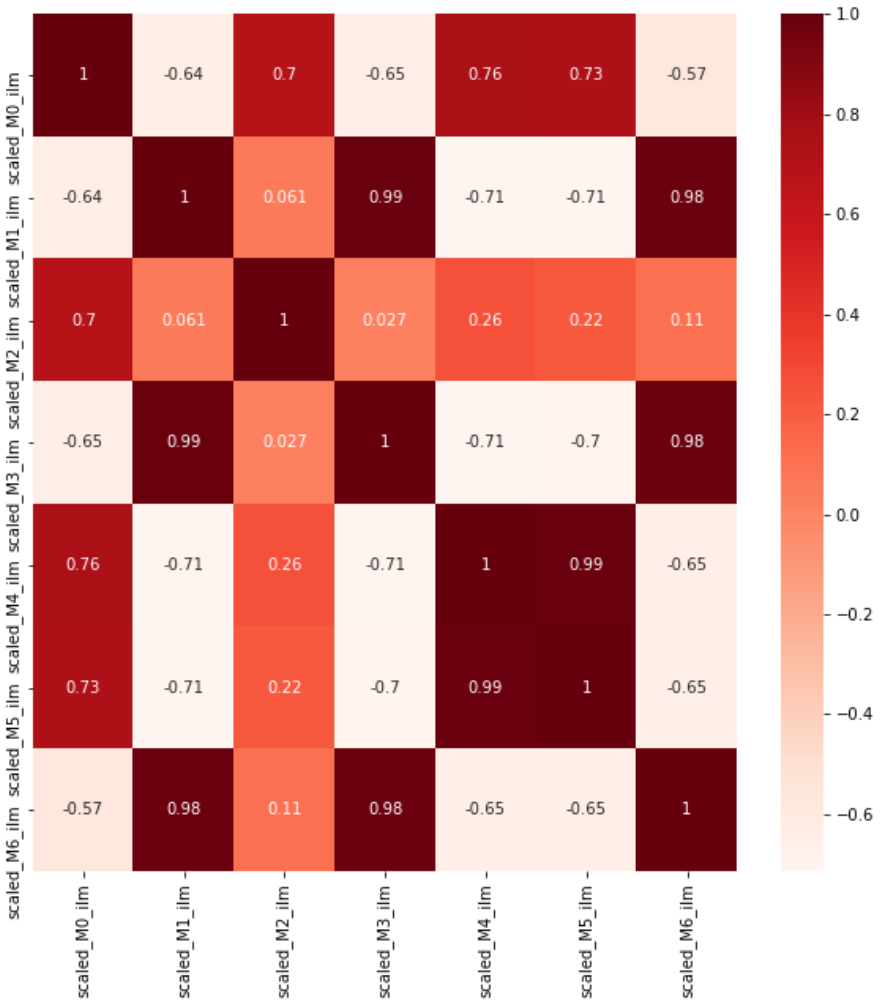


Figura 3.5: Coeficiente de correlación de Pearson de los momentos sobre toda la ILM

de Pearson.

Para poder afirmar las correlaciones entre los distintos momentos geométricos, se estudia también la matriz de correlación de Pearson de los distintos invariantes describiendo toda la ILM. Así obtenemos el mapa de calor que se observa en la figura 3.5. Como puede verse, se mantienen las correlaciones más evidentes descritas anteriormente. Sin embargo, el coeficiente de correlación disminuye en el caso del primer momento y el tercero. Por lo tanto, es posible que no se trate de una correlación tan robusta como la del resto de parejas de invariantes mencionadas. Asimismo en dicha gráfica comienzan a apreciarse indicios de correlaciones lineales negativas que no existían cuando los momentos se restringían a la cúpula. También, cabe volver a recordar que el conjunto de invariantes aplicados a toda la ILM posee un mayor número de elementos, característica que hace más representativa esta última figura.

### 3.4.2. Matriz de dispersión

Una representación gráfica con la que es posible tanto visualizar las correlaciones como tratar de encontrar diferencias significativas entre el conjunto de pacientes glaucomatosos y el grupo de control es la matriz de dispersión. Esta herramienta de exploración de datos representa los distintos elementos del sistema tomando dos dimensiones, de manera que se genera una matriz en la que cada celda es el resultado de proyectar cada vector sobre el plano conformado por los momentos en cuestión.

Si se genera la matriz de dispersión a partir de los momentos geométricos restringidos a la excavación se obtiene la figura 3.6. En ella los elementos de cada uno de los dos conjuntos se distinguen porque han sido representados en colores distintos (rojo y azul). Lo que más destaca en la gráfica son las correlaciones que salen a la luz en la sección 3.4.1 que pueden verse claramente. Sin embargo, no resulta evidente encontrar una proyección en la que sea sencillo distinguir los pacientes glaucomatosos del grupo de control.

En la figura 3.7, se puede observar la matriz de dispersión de los momentos geométricos aplicados sobre toda la ILM. En ella, al ser un conjunto con un mayor número de elementos, se hace más patente las correlaciones entre los distintos momentos. Asimismo, al igual que ocurre restringiendo los invariantes a la excavación, tampoco se diferencian claramente los dos conjuntos en ninguna de las celdas presentes de la matriz.

A pesar de no haber obtenido nuevas conclusiones, estas representaciones gráficas son de utilidad y permiten extraer conclusiones con técnicas de aprendizaje automático como el análisis de componentes principales (véase 4.2).

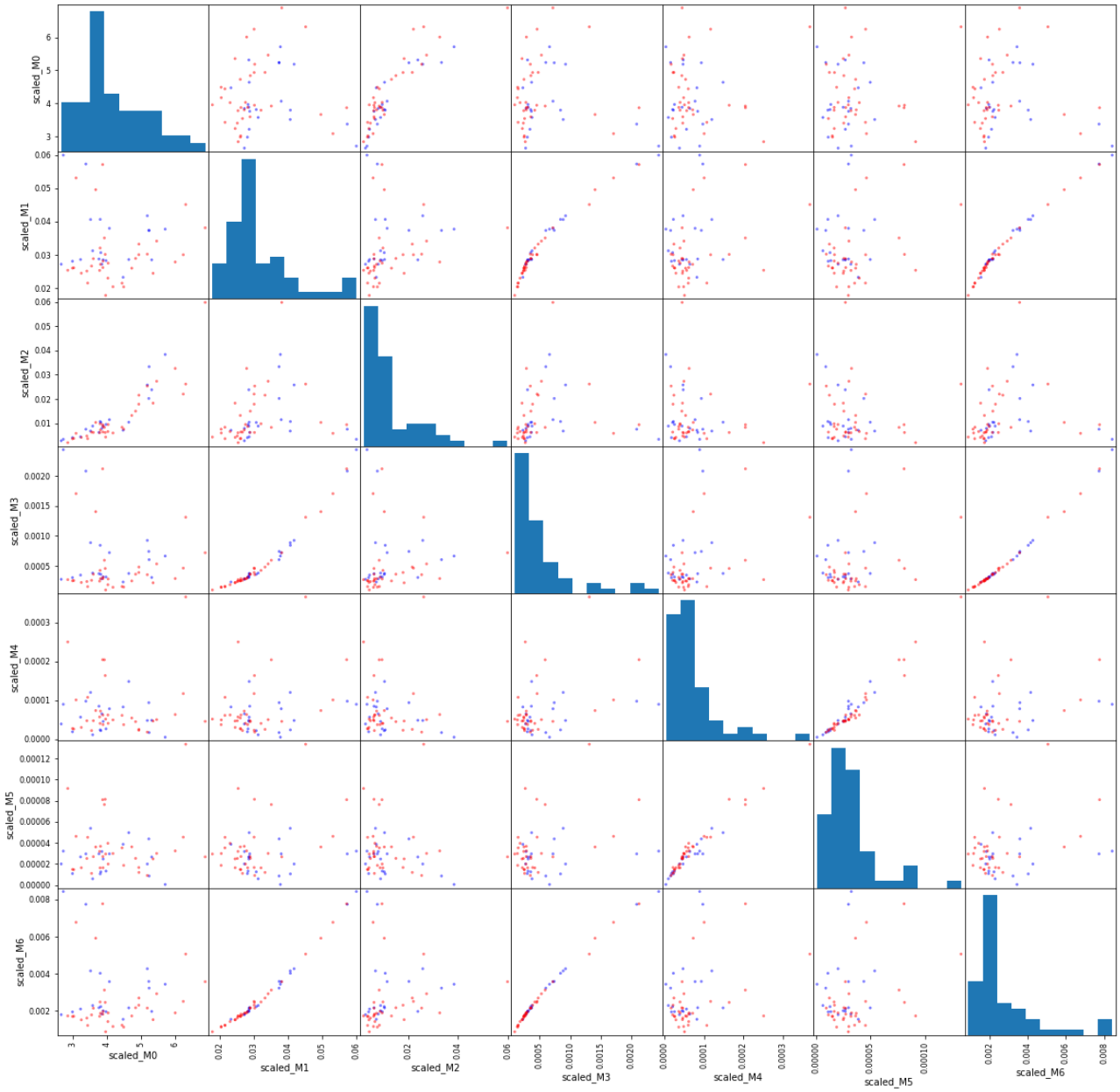


Figura 3.6: Matriz de dispersión de los momentos geométricos restringidos a la excavación

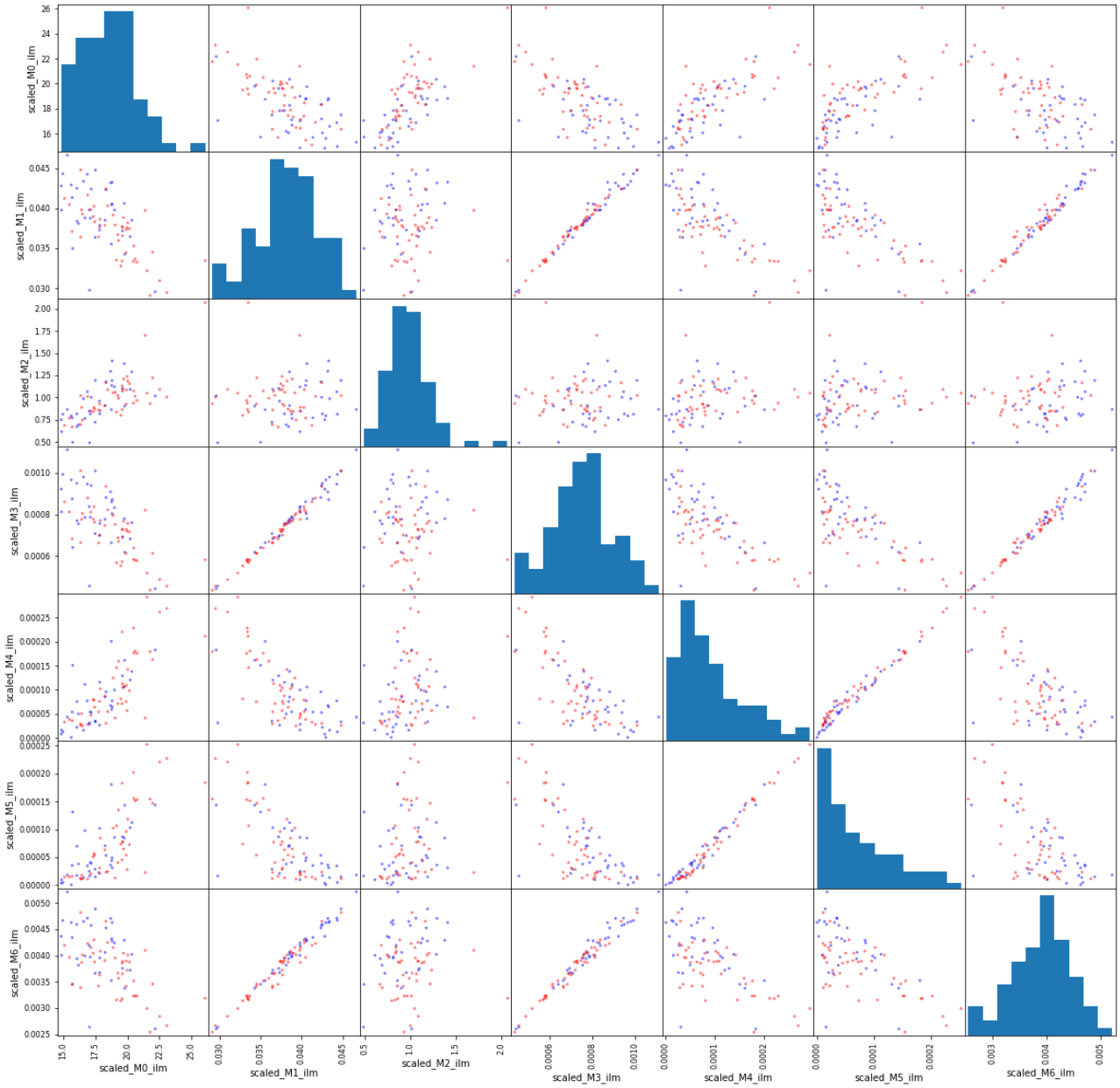


Figura 3.7: Matriz de dispersión de los momentos geométricos de toda la ILM

Capítulo 4

# Análisis de datos con técnicas de machine learning

*"Si torturas lo suficiente a tus datos, estos te confesarán lo que quieras"*
— Ronald H. Coase (Premio Nobel de Economía en 1991)

En este capítulo se utilizan técnicas de aprendizaje automático para sacar conclusiones del sistema de datos. Se comienza con una análisis preliminar utilizando algoritmo de clustering (véase la sección 4.1) para comprobar cómo de bien se ajustan los distintos momentos geométricos invariantes cuando se trata de automatizar la tarea de diagnóstico. A continuación, con el fin de desvelar características generales del conjunto de datos, se lleva a cabo un análisis de componentes principales con el que se obtendrán fuertes dependencias del sistema sobre ciertos invariantes (véase la sección 4.2). Por último, se utilizan los árboles de decisión tanto para construir un modelo de predicción como para extraer conclusiones que describan el conjunto de los datos (véase la sección 4.3).

## 4.1. Análisis preliminar de los momentos utilizando algoritmos de clustering

En una primera aproximación, es interesante conocer cómo de bien se ajustan los momentos geométricos a la hora de discernir entre un paciente que padece glaucoma y otro que pertenece al grupo de control del estudio. Es decir, nos preguntamos si los momentos geométricos (de maneral individual o conjunta con otros) determinan el diagnóstico del glaucoma de forma directa. Para medir esta calidad en la tarea de diferenciación, haremos uso de dos algoritmos de clutering (Xu y Wunsch, 2008) muy utilizados, los cuales agruparán de manera automática el conjunto de elementos (cada elemento es una 7-tupla cuyas componentes vienen determinadas por los distintos momentos geométricos) de manera que los miembros de un mismo grupo (denominado cúmulo o *cluster*) son más "parecidos" entre ellos en función de un criterio concreto (por ejemplo es posible definir la similitud entre elementos como la distancia euclídea presente entre los puntos en el espacio 7-dimensional). Los algoritmos que utilizaremos serán los llamados *K-Medias* (Hartigan, 1975) y *DBSCAN* (Ester et al., 1996), cuyas siglas hacen referencia a su nombre en inglés *Density-Based Spatial Clustering of Applications with Noise*. Si consiguiéramos un buen resultado con alguno de estos algoritmos, podríamos diagnosticar la enfermedad de forma automática identifi-

cando el cluster (grupo de enfermos o de control) al que pertenece el paciente objetivo.

Aunque ambos algoritmos requieran un parámetro inicial (en el caso de K-Medias el parámetro es el número $k$ de clusters que deben generarse y, en el caso de DBSCAN, el parámetro es la distancia umbral $\varepsilon$ que define cada vecindad o cúmulo) que debe establecerse antes de la ejecución, solo será necesario estudiar la variación del correspondiente al algoritmo DBSCAN. Esto se debe a que, dada la naturaleza del enunciado del problema de clasificación que queremos abordar, conocemos de antemano el número $k$ de clusters que queremos que se generen: 2 (pacientes sanos y afectados por el glaucoma). A pesar de ello, evaluaremos "medidas de ayuda a la decisión" como el coeficiente de Silhouette (Rousseeuw, 1987) sobre ambos algoritmos, ya que estas nos permiten valorar la dispersión interna y la presente entre los clusters generados.

### 4.1.1. El coeficiente de Silhouette

Una de las medidas de ayuda a la decisión más famosas es el coeficiente de Silhouette (SC). Esta técnica de interpretación y validación de la consistencia propia de cada cluster, proporciona un valor que mide la similitud de un elemento con el cluster al que pertenece frente al resto de cúmulos. Antes de definirlo formalmente, es necesario entender algunos conceptos previos que explicamos a continuación.

Entendamos los elementos de un sistema $S$, que son $n$-tuplas, como elementos del espacio euclídeo $\mathbb{R}^n$ (por lo que $S \subseteq \mathbb{R}$). Sean $a = (a_1, \ldots, a_n), b = (b_1, \ldots, b_n) \in S$ y $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$, entonces podemos definir la *p-distancia* con peso $w$ entre ellos como:

$$d_p(a,b) := \|a-b\|_p := \left( \sum_{i=1}^{n} w_i |a_i - b_i|^p \right)^{1/p}$$

El peso $w_i$ representa la "importancia" del $i$-estado (la $i$-ésima dimensión del sistema), pudiendo ser $w$ equidistribuido, es decir, $w_i = 1 \ \forall i \in \{1, \ldots, n\}$.

Con el concepto de p-distancia, podemos definir la distancia media de un elemento a su cluster. Definimos cluster como un subconjunto no vacío del sistema. Sean $C \subseteq S$ un cúmulo (y denotaremos por $|C|$ el número de elementos) con al menos un elemento distinto a $a \in C$. Entonces definimos la *p-distancia media* de $a$ a su cluster como:

$$\overline{d}_p(a) := \frac{1}{|C|-1} \sum_{b \in C} d_p(a,b)$$

Otro concepto que necesitaremos es la p-distancia mínima de $a \in S$ al resto de cúmulos. Sean $S \subset \mathbb{R}^n$ un sistema de $N \in \mathbb{N}$ elementos distribuidos en $k > 1$ clusters (que por definición son no vacíos) $\{C_q\}_{q=1}^k$ y $a \in C_{q_0}$ tal que $q_0 \in \{1, \ldots, k\}$. De esta forma definimos la *p-distancia mínima* de $a$ al resto de cúmulos como:

$$d_p(a, \{C_q\}_{q=1}^k) := \text{mín} \left\{ \frac{1}{|C_q|} \sum_{b \in C_q} d_p(a,b) : q \in \{1, \ldots, k\} \setminus \{q_0\} \right\}$$

Con estos conceptos previos, ya podemos definir el coeficiente de Silhouette. Sean $S \subset \mathbb{R}^n$ un sistema de $N \in \mathbb{N}$ elementos distribuidos en $k > 1$ clusters $\{C_q\}_{q=1}^k$ y $a \in C_{q_0}$ tal

que $q_0 \in \{1, \ldots, k\}$. Definimos la silueta o *silhouette* de $a$ como:

$$\overline{s}_p(a) := \begin{cases} \dfrac{d_p(a, \{C_q\}_{q=1}^k) - \overline{d}_p(a)}{\max\{\overline{d}_p(a), d_p(a, \{C_q\}_{q=1}^k)\}} & \text{si } |C_{q_0}| > 1 \\ 0 & \text{si } |C_{q_0}| = 1 \end{cases}$$

Lo cual es equivalente a escribir:

$$\overline{s}_p(a) := \begin{cases} 1 - \dfrac{\overline{d}_p(a)}{d_p(a, \{C_q\}_{q=1}^k)} & \text{si } \overline{d}_p(a) < d_p(a, \{C_q\}_{q=1}^k) \\ 0 & \text{si } \overline{d}_p(a) = d_p(a, \{C_q\}_{q=1}^k) \\ \dfrac{d_p(a, \{C_q\}_{q=1}^k)}{\overline{d}_p(a)} - 1 & \text{si } \overline{d}_p(a) > d_p(a, \{C_q\}_{q=1}^k) \end{cases}$$

Por lo tanto, resulta evidente que $\overline{s}_p(a) \in (-1, 1)$. Una observación interesante es que para que $\overline{s}_p(a)$ esté cerca del valor 1, debe tenerse que $\overline{d}_p(a) \ll d_p(a, \{C_q\}_{q=1}^k)$. Como $\overline{d}_p(a)$ es una medida de la disimilitud del elemento $a$ con su cúmulo, cuanto menor sea el valor de $\overline{d}_p(a)$ más "se parecerá" (en términos de p-distancia) $a$ al resto de su cluster. Por otro lado, cuanto mayor sea el valor de $d_p(a, \{C_q\}_{q=1}^k)$ significará que $a$ es "más distinto" de los elementos agrupados en otros cúmulos. Por lo tanto, cuanto más cerca se encuentre $\overline{s}_p(a)$ de 1 podremos afirmar que la agrupación de $a$ es "mejor".

Por el contrario, como cabía esperar, para que $\overline{s}_p(a)$ esté cerca del valor -1, debe cumplirse que $\overline{d}_p(a) \gg d_p(a, \{C_q\}_{q=1}^k)$. Utilizando los mismos argumentos que en el caso positivo se deduce que cuanto más cerca esté $\overline{s}_p(a)$ de -1 más indicativo será de que el elemento $a$ debería clasificarse en un cluster distinto.

Por último, siguiendo el razonamiento anterior, resulta sencillo deducir que si $\overline{s}_p(a)$ es un valor cercano a 0 podemos afirmar que el elemento $a$ se encuentra prácticamente en una posición de equidistancia entre dos cúmulos distintos.

Definiremos el *coeficiente de Silhouette* como la media de $\overline{s}_p(a)$ sobre todos los elementos de nuestro sistema $S$, es decir, suponiendo que $\forall q \in \{1, \ldots, k\}$ $|C_q| > 1$ tendremos:

$$\overline{s}_p(\{C_q\}_{q=1}^k) := \frac{1}{N} \sum_{q=1}^{k} \sum_{a \in C_q} \frac{d_p(a, \{C_q\}_{q=1}^k) - \overline{d}_p(a)}{\max\{\overline{d}_p(a), d_p(a, \{C_q\}_{q=1}^k)\}}$$

Intuitivamente, es posible deducir que la media de todas las siluetas de los elementos de un cúmulo nos indica cómo de bien agrupado está dicho cluster. Por lo tanto el coeficiente de Silhouette, que es precisamente la media de todas las siluetas de todos los elementos del sistema, será una medida de la calidad y eficacia (en términos de p-distancia) con las que han sido generados todos los cúmulos.

Podemos encontrar otras concepciones distintas del coeficiente de Silhouette, como la propuesta por Kaufman y Rousseeuw (2009) en la que lo define como:

$$SC := \max\{\overline{s}_p(\{C_q\}_{q=1}^k) : k \in \mathbb{N}\}$$

De manera que lo concibe como el máximo valor que se puede obtener con $\overline{s}_p(\{C_q\}_{q=1}^k)$ entre todos los posibles números $k$ de cúmulos generados. Sin embargo, utilizaremos la definición presentada originalmente (en lugar de esta propuesta por Kaufman y Rousseeuw (2009) o la de otros autores), ya que es la implementada en la librería de machine learning de Python *sklearn*.

### 4.1.2. Índice de Rand ajustado

Además de conocer la calidad de nuestra distribución en cúmulos, estamos interesados en medir las similitudes de los clusters resultados de la ejecución con la clasificación real de los elementos entre controles y pacientes de glaucoma. De esta forma podremos cuantificar el ajuste que se obtiene tras la finalización de cada algoritmo de clustering.

Para abordar este problema utilizamos otra famosa medida de ayuda a la decisión conocida como el índice de Rand ajustado (ARI), el cual es una forma del Índice de Rand, propuesto por Rand (1971), que se ajusta a la agrupación aleatoria de los elementos. Por esta razón, definiremos en primer lugar este último para dar, luego, lugar a la explicación de la medida que utilizaremos.

Sea $S \subset \mathbb{R}^n$ un sistema de $N \in \mathbb{N}$ elementos distribuidos en dos conjuntos de cúmulos $\{C_q^1\}_{q=1}^{k_1}$ y $\{C_q^2\}_{q=1}^{k_2}$ con $k_1$ y $k_2$ clusters, respectivamente. Definiremos los coeficientes de Rand como:

- $r_1 := |\{(a,b) : \exists q_1, q_2 \text{ tales que } a, b \in C_{q_1}^1 \cap C_{q_2}^2\}|$, es decir, $r_1$ es el número de pares de elementos de $S$ que se encuentran en un mismo cúmulo tanto en la distribución $\{C_q^1\}_{q=1}^{k_1}$ como en la $\{C_q^2\}_{q=1}^{k_2}$.
- $r_2 := |\{(a,b) : \exists q_{11}, q_{12}, q_{21}, q_{22}\text{que cumplen que } q_{11} \neq q_{12} \wedge q_{21} \neq q_{22} \text{ tales que se tiene } a \in C_{q_{11}}^1 \cap C_{q_{21}}^2 \wedge b \in C_{q_{12}}^1 \cap C_{q_{22}}^2\}|$, es decir, $r_2$ es el número de pares de elementos de $S$ que se encuentran en distintos cúmulos en ambas distribuciones.
- $r_3 := |\{(a,b) : \exists q_{11}, q_{12}, q_2 \; q_{11} \neq q_{12} \text{ tal que } a \in C_{q_{11}}^1 \cap C_{q_2}^2 \wedge b \in C_{q_{12}}^1 \cap C_{q_2}^2\}|$, es decir, es el número de pares de elementos del sistema que en la primera distribución no se encuentran en un mismo cúmulo y en la segunda sí.
- $r_4 := |\{(a,b) : \exists q_1, q_{21}, q_{22} \; q_{21} \neq q_{22} \text{ tal que } a \in C_{q_1}^1 \cap C_{q_{21}}^2 \wedge b \in C_{q_1}^1 \cap C_{q_{22}}^2\}|$, es decir, es el número de pares de elementos del sistema que en la primera distribución se encuentran en un mismo cúmulo y en la segunda no.

Intuitivamente, es posible interpretar $r_1 + r_2$ como el número de pares coincidentes entre las dos agrupaciones en cúmulos y $r_3 + r_4$ como el número de pares distribuidos de manera distinta en $\{C_q^1\}_{q=1}^{k_1}$ y $\{C_q^2\}_{q=1}^{k_2}$. De esta manera, definimos el *índice de Rand* como:

$$RI := \frac{r_1 + r_2}{r_1 + r_2 + r_3 + r_4} = \frac{r_1 + r_2}{\binom{N}{2}} = \frac{r_1 + r_2}{\frac{N}{2}(N-1)}$$

Como el denominador es el número total de pares que es posible formar con nuestro sistema $S$, el índice de Rand nos indicará la probabilidad de encontrarnos con una coincidencia entre ambas distribuciones. Por ello, es fácil ver que tomará valores entre 0 y 1 como otra función de probabilidad.

El entendimiento de la interpretación del índice de Rand facilita la comprensión del índice de Rand ajustado. Este último es la corrección mediante el uso del modelo de permutación del $RI$ (Wagner y Wagner, 2007), de manera que se define como:

$$\frac{RI - RIEsperado}{\text{máx}(RI) - RIEsperado}$$

Para facilitar el cálculo de cada elemento de la definición, se introduce el concepto de *coeficientes de Rand ajustados* (Gates y Ahn, 2017). Sean $i \in \{1, \ldots, k_1\}$ y $j \in \{1, \ldots, k_2\}$, entonces definimos el coeficiente $ij$ de Rand ajustado como $n_{ij} := |C_i^1 \cap C_j^2|$, es decir, el número de elementos comunes en los cúmulos $C_i^1$ y $C_j^2$ de las dos distribuciones.

De esta manera, podemos calcular el índice de Rand ajustado como:

$$\frac{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}\binom{n_{ij}}{2} - \left[\sum_{i=1}^{k_1}\binom{a_i}{2}\sum_{j=1}^{k_2}\binom{b_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_{i=1}^{k_1}\binom{a_i}{2} + \sum_{j=1}^{k_2}\binom{b_j}{2}\right] - \left[\sum_{i=1}^{k_1}\binom{a_i}{2}\sum_{j=1}^{k_2}\binom{b_j}{2}\right]/\binom{N}{2}}$$

donde $a_i = |C_i^1|$ y $b_j = |C_j^2|$. Con esta definición, podemos afirmar que el índice de Rand ajustado podría tomar valores negativos cuando el índice de Rand es menor que el esperado.

### 4.1.3. K-Medias

Ahora que se han explicado las medidas de ayuda a la decisión con las que evaluaremos las clasificaciones obtenidas (el coeficiente de Silhouette, explicado en la sección 4.1.1, y el índice de Rand ajustado, explicado en la sección 4.1.2), se estudia el comportamiento de uno de los dos algoritmos de clustering que se ejecutan: K-Medias (Hartigan, 1975). Este método de agrupación divide el espacio en celdas de Voronoi y clasifica los elementos en ellas. Como se comentó anteriormente, este algoritmo requiere como parámetro inicial $k$ el número de cúmulos en los que deben agruparse los elementos del sistema. En nuestro caso, el valor debe ser $k = 2$ debido a que estamos interesados en detectar cuándo un paciente padece glaucoma (o por el contrario es una persona sana).

El algoritmo K-Medias también requiere que se elija el método con el cual se obtendrán los centroides que inicializarán el proceso. Existen diversas formas de abordar este problema, sin embargo en este estudio se utiliza el *método de Forgy* (Hamerly y Elkan, 2002), también conocido como método aleatorio, que es el más empleado. Para inicializar el algoritmo, escogemos al azar $k$ elementos del sistema que serán los centroides de la primera iteración.

Dado que la aleatoriedad intervendrá y podría ser determinante en el resultado final del algoritmo, se ejecuta un número notable de veces (concretamente 1000 veces) para así conocer la distribución aleatoria tanto del coeficiente de Silhouette como del índice de Rand ajustado y obtener conclusiones a partir de ella. De esta manera, aplicando el algoritmo sobre los datos de los momentos geométricos restringidos a la excavación, obtenemos como resultado las distribuciones del coeficiente de Silhouette de la figura 4.1 y del índice de Rand ajustado de la figura 4.2.

Analizando, en primer lugar, la distribución del coeficiente de Silhouette (figura 4.1) tras ejecutar 1000 veces el algoritmo K-Medias sobre los momentos geométricos restringidos

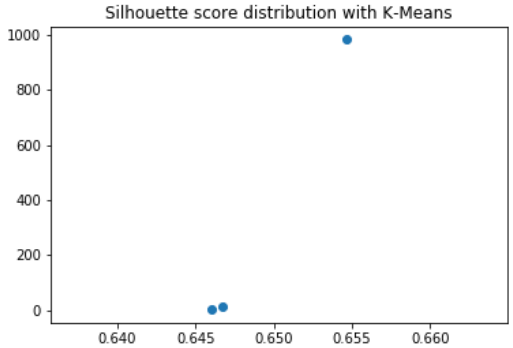


Figura 4.1: Distribución del SC (K-Medias)

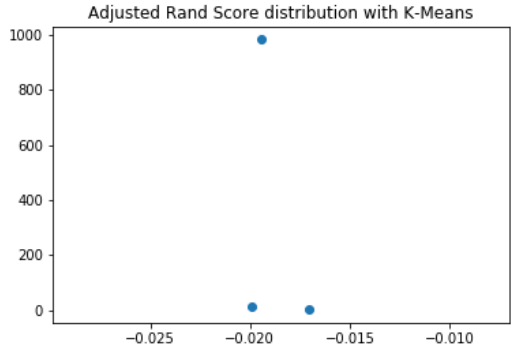


Figura 4.2: Distribución del ARI (K-Medias)

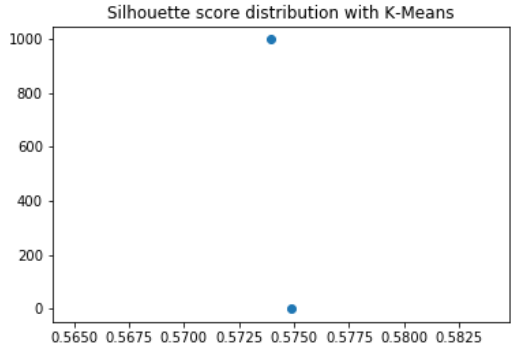


Figura 4.3: Distribución del SC (K-Medias)

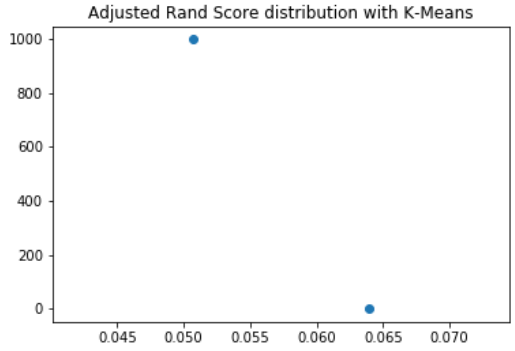


Figura 4.4: Distribución del ARI (K-Medias)

a la excavación de la imagen SLO, se encuentra un valor del coeficiente (aproximadamente 0,655) que se repite un número mayor de veces (concretamente 983 veces, lo que representa el 98,3 % de la muestra). La diferencia con el resto de posibles valores (que se encuentran alrededor de 0,6467 y 0,646), en lo que se refiere a frecuencia absoluta (los otros valores aparecen tan solo en 14 y 3 ocasiones, respectivamente), resulta suficientemente notable como para poder afirmar que el coeficiente de Silhouette de nuestro sistema al ejecutar el algoritmo K-Medias es de 0,646. Como se ha explicado en la sección 4.1.1, este valor nos indica que, en general, la p-distancia mínima de los elementos a otros cúmulos es considerablemente mayor que la p-distancia media a sus propios clusters. Por ello, nos encontramos ante una clasificación bastante consistente.

En lo que se refiere a la distribución del índice de Rand ajustado (figura 4.2), al igual que ocurre con el coeficiente de Silhouette, es posible observar un valor (aproximadamente $-0{,}0194$), en este caso negativo, que se repite un mayor número de veces (concretamente 983 veces, al igual que el valor preponderante del coeficiente de Silhouette). También se han obtenido dos valores distintos (aproximadamente $-0{,}0199$ y $-0{,}017$) que aparecen en la muestra en un número de ocasiones notablemente menor (en 14 y 3 ocasiones). Sin embargo, haciendo uso de lo visto en la sección 4.1.2, estos resultados nos indican que las clasificaciones obtenidas por el algoritmo, a pesar de ser consistentes, divergen mucho de la categorización real entre pacientes sanos y que padecen glaucoma. Por lo tanto, podemos concluir que no es posible, con los datos con los que se cuenta, diagnosticar de manera automática la enfermedad a partir de las celdas de Voronoi generadas por el algoritmo K-Medias, ya que la tasa de error es increíblemente alta dada la distribución obtenida del índice de Rand ajustado.

Ante las conclusiones obtenidas, se plantea la cuestión de si conseguiríamos resultados similares si no se restringieran los momentos geométricos a la excavación y se analizara toda la ILM (caso con el que se cuenta con un mayor número de elementos en el sistema). Por este motivo, seguimos el mismo procedimiento que el explicado anteriormente ejecutando el algoritmo K-Medias 1000 veces y extrayendo de estas iteraciones la distribución tanto del Coeficiente de Silhouette como del índice de Rand ajustado. Las distribuciones obtenidas como resultado de dicha ejecución son las que se muestran en las figuras 4.3 y 4.4, respectivamente.

En esta ocasión se obtiene una muestra más simple (de tan solo dos valores) del coeficiente de Silhouette (véase la figura 4.3), en la que encontramos dos frecuencias absolutas completamente extremas: 999 para un valor aproximado de 0,574 y 1 para un coeficiente aproximado de 0,575. Este resultado es ligeramente inferior al extraído restringiendo los

momentos geométricos a la excavación, pero la diferencia es despreciable, ya que de ambos coeficientes de Silhouette podemos deducir una relevante consistencia en la clasificación resultante de la ejecución del algoritmo de K-Medias.

En cuanto a la distribución del índice de Rand ajustado (figura 4.4), también se observan dos valores distintos (aproximadamente de 0,051 y 0,064) con frecuencias extremas (999 y 1, respectivamente). A pesar de obtener resultados superiores a los mostrados al restringir los momentos geométricos, la diferencia con estos vuelve a ser despreciable, ya que el índice de Rand ajustado es excesivamente bajo como para poder clasificar automáticamente (con una precisión considerable) a través de los cúmulos generados por el algoritmo.

Tras obtener unos valores del índice de Rand ajustado considerablemente bajos (tanto restringiendo los momentos geométricos como analizando toda la ILM), se debe descartar la automatización, mediante el algoritmo de K-Medias, del diagnóstico de glaucoma. Por este motivo, a continuación se estudian la misma cuestión para otro algoritmo de clustering que presenta un comportamiento distinto.

### 4.1.4. DBSCAN

El algoritmo DBSCAN (Ester et al., 1996) es un método de agrupamiento cuyo concepto principal se basa en la localización de regiones de alta densidad que se encuentren separadas entre sí por zonas de baja densidad (esta idea general también subyace en otros algoritmos como el llamado OPTICS (Ankerst et al., 1999), cuyas siglas provienen de su denominación en inglés: *Ordering Points To Identify the Clustering Structure*). Para llevar esta concepción a la práctica, se define la vecindad o $\varepsilon$-bola $V_a^\varepsilon$ de un elemento $a \in S$ del sistema dada por una p-distancia como:

$$V_a^\varepsilon = \{b \in S : d_p(a, b) \leq \varepsilon\}$$

De esta manera se pueden generar los distintos cúmulos atendiendo a esa p-distancia umbral dada por $\varepsilon$. Por lo tanto, como se ha explicado, este algoritmo requiere la inicialización de dicha p-distancia previo a su ejecución.

Como se desconoce la p-distancia umbral con la que se podrán obtener unos resultados adecuados para el objetivo de automatización del diagnóstico, se ejecutará el algoritmo DBSCAN variando dicho parámetro $\varepsilon$ y, para cada clasificación en cúmulos, se evaluará el coeficiente de Silhouette y el índice de Rand ajustado para, de esta manera, poder extraer conclusiones en función de la variable de inicialización. Además, como la implementación de este método en la librería de Python *sklearn* permite la elección de distintas distancias para su ejecución, se extraerán resultados utilizando tanto la distancia euclídea (o 2-distancia) y la distancia Manhattan.

Comenzando por el análisis de la ejecución del algoritmo DBSCAN (con la distancia euclídea) sobre los momentos geométricos restringidos a la excavación, se obtienen las gráficas que pueden observarse en la figura 4.5. En la primera se muestra el coeficiente de Silhouette obtenido en función de la distancia umbral elegida. Como puede observarse, alrededor de un valor de $\varepsilon$ de 0,2 se consiguen valores razonables que indican una notable consistencia de la distribución en cúmulos generada. La segunda gráfica de la figura 4.5 relaciona la variación del parámetro de inicialización con el número de clusters obtenidos finalmente. Alrededor del valor de 0,2 se obtiene, en general, 3 cúmulos resultantes en la clasificación, lo cual puede ser razonable y no difiere en exceso de la categorización real que se quiere obtener. En ambas representaciones gráficas se marca el valor del parámetro $\varepsilon$ con el que se alcanza el máximo del coeficiente de Silhouette en el intervalo evaluado (los puntos

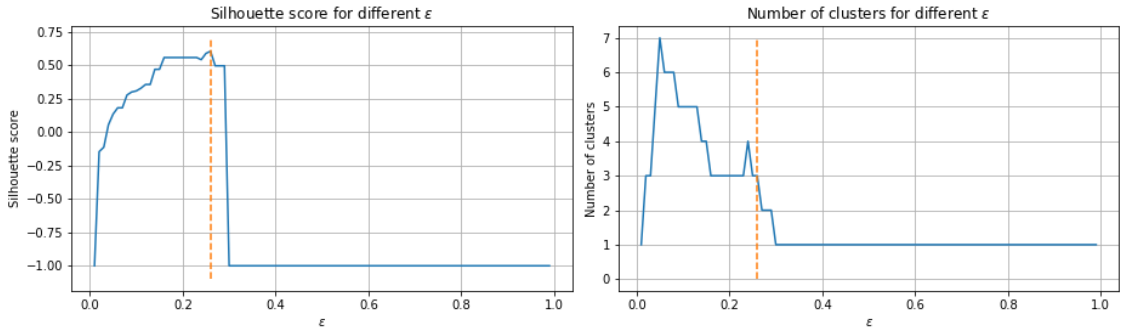


Figura 4.5: Coeficiente de Silhouette con la distancia euclídea (DBSCAN)

tomados son aquellos que pertenecen al conjunto $\{0{,}01+n\cdot 0{,}01 : n \in \mathbb{Z} \wedge 0 \leq k < 100\}$). Un hecho reseñable es la repentina caída del coeficiente de Silhouette en un punto ligeramente superior al valor de $\varepsilon$ donde se alcanza el máximo, la cual va acompañada también de una disminución en el número de cúmulos clasificados (obteniendo tan solo un cluster). Gracias a esta observación podemos concluir que a partir de dicha distancia umbral la agrupación deja de ser suficientemente consistente.

La misma ejecución del algoritmo DBSCAN sobre los momentos geométricos restringidos a la excavación y con el mismo conjunto que recorre el parámetro de inicialización, pero variando la distancia utilizada para delimitar las vecindades (en esta ocasión se utiliza la distancia manhattan), saca a la luz resultados muy similares a los anteriores. Este hecho puede observarse en las dos gráficas de la figura 4.6. Si se comparan con las de la figura 4.5, se puede afirmar que los valores máximos del coeficiente de Silhouette son muy parecidos, así como el número de cúmulos que han generado (en la mayoría de casos 3 clusters). Además también comparten esa repentina disminución tanto del coeficiente de Silhouette como del número de clusters resultantes cerca del máximo. Por lo tanto, en este caso, no influirá de manera significativa la distancia elegida (entre las dos que se han evaluado) a la hora de diagnosticar la enfermedad, ya que parece que se obtienen resultados muy similares.

A pesar de la demostrada consistencia de los cúmulos resultantes de la clasificación, ambas distancias presentan un índice de Rand ajustado excesivamente bajo como para poder automatizar la tarea de diagnóstico. Como vemos en las figuras 4.7 y 4.8, no se encuentra apenas similitud alguna entre la clasificación real y la obtenida tras la ejecución del algoritmo DBSCAN. De hecho, an ambos casos, la mayor parte de los valores son nulos o negativos, lo cual muestra la poca coincidencia que hay entre las distribuciones.

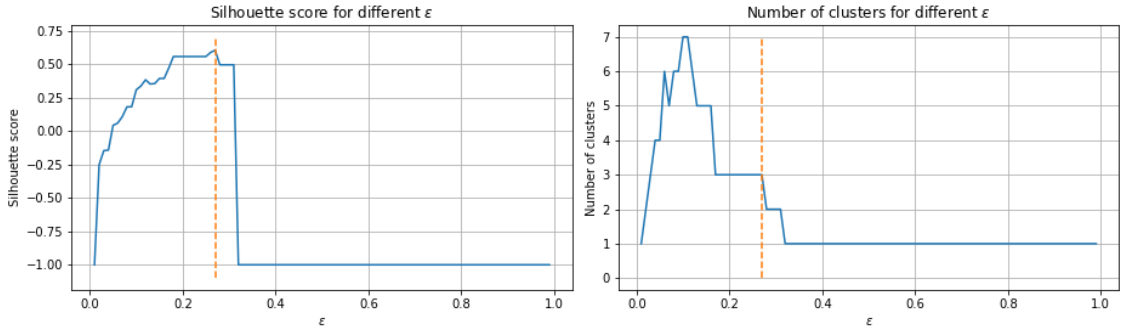


Figura 4.6: Coeficiente de Silhouette con la distancia manhattan (DBSCAN)

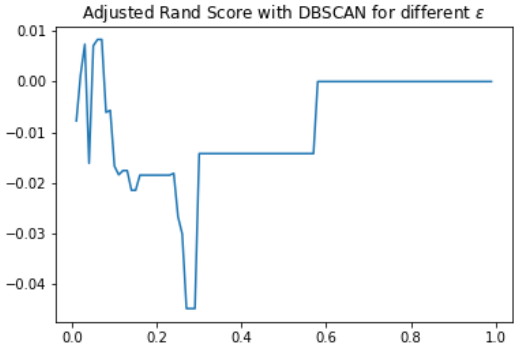


Figura 4.7: ARI 2-distancia (DBSCAN)

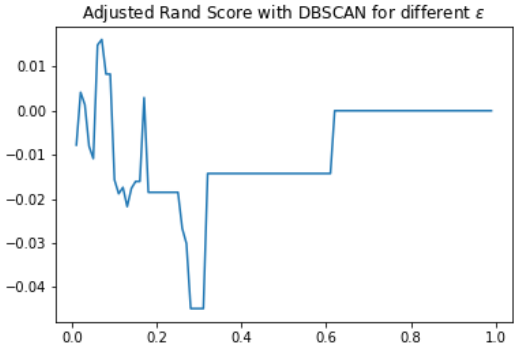


Figura 4.8: ARI manhattan (DBSCAN)

Una vez más, las gráficas de las distintas distancias vuelven a parecerse, lo cual indica que, con bastante certeza, se obtienen distribuciones en cúmulos similares, es decir, que la clasificación es independiente de si se escoge la distancia euclídea o la distancia manhattan.

Ante la poca fiabilidad de automatizar el diagnóstico, se analiza la viabilidad de llevarlo a cabo con el algoritmo DBSCAN sobre los momentos geométricos de toda la ILM (en lugar de restringirlos a la excavación). Asimismo se tiene la oportunidad de llevar a cabo el estudio con un mayor número de datos y extraer conclusiones más representativas debido al mayor tamaño de la muestra.

La evolución del coeficiente de Silhouette respecto de la distancia umbral se puede observar en la primera gráfica de la figura 4.9. En ella se puede ver como, a pesar de obtener valores ligeramente inferiores a los resultantes en el caso de los momentos geométricos restringidos a la excavación, hay un mayor número de parámetros de inicialización con los que se obtienen cúmulos consistentes. Por otro lado, se obtiene también un número de clusters que concuerda mejor con la clasificación real (en gran parte del intervalo en el que $\varepsilon$ toma valores que generan agrupaciones consistentes, se generan dos cúmulos), ya que, como se menciona anteriormente, en el caso previo se conseguían 3 clusters.

Utilizando la distancia manhattan, en esta ocasión (probablemente se deba a que se cuenta con un mayor número de elementos en el sistema), si es posible visualizar pequeñas diferencias en el comportamiento del coeficiente de Silhouette en comparación con el obtenido al usar la distancia euclídea. Si se observa la gráfica de la izquierda de la figura 4.10, se pueden vislumbrar algunos aspectos en los que ambas difieren. Sin entrar a analizar la existencia de un mínimo en un punto aislado que no se presenta en la primera gráfica de la

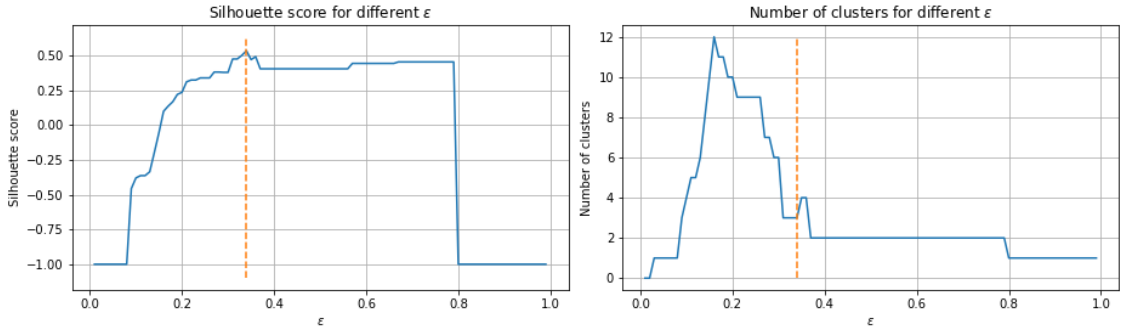


Figura 4.9: Coeficiente de Silhouette con la distancia euclídea (DBSCAN)

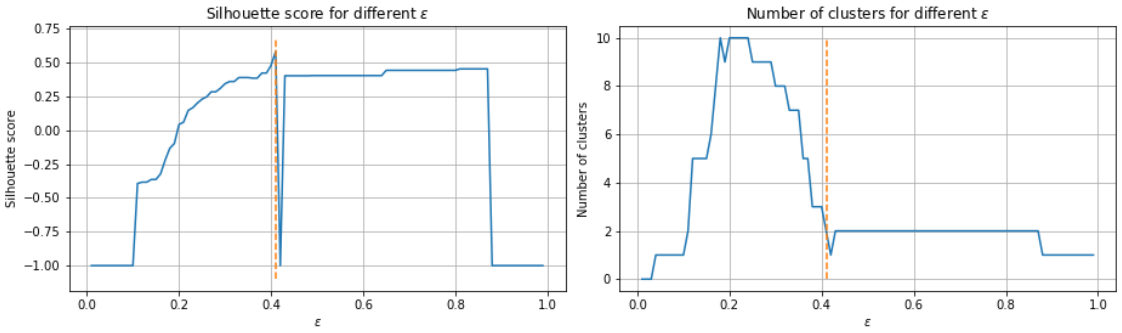


Figura 4.10: Coeficiente de Silhouette con la distancia manhattan (DBSCAN)

figura 4.9, resulta notable que el crecimiento del coeficiente de Silhouette en el caso de la distancia manhattan es menos pronunciado y requiere de valores de $\varepsilon$ mayores para alcanzar los mismos puntos que en el primer caso. Sin embargo, con la gráfica de la figura 4.10 se llegan a valores ligeramente superiores a los coeficientes de Silhouette obtenidos con la distancia euclídea. Asimismo, en contraposición con su crecimiento "más lento", la gráfica generada al utilizar la distancia manhattan, genera distribuciones de cúmulos consistentes con valores mayores del parámetro de inicialización (como puede observarse el coeficiente de Silhouette alcanza valores negativos cerca de 0,9 de distancia umbral, mientras que en el caso de la distancia euclídea desciende antes del valor de 0,8 en el parámetro). Estas pequeñas diferencias también afectan al número de cúmulos resultantes que, aunque en ambos casos se obtienen, en general, dos, con la distancia manhattan se mantienen dos clusters con valores superiores de $\varepsilon$.

No obstante, pese a la generación de cúmulos consistentes utilizando ambas distancias, en el caso de los momentos geométricos de toda la ILM, se vuelven a observar valores excesivamente bajos del índice de Rand ajustado (esto puede verse en las figuras 4.11 y 4.12). Esto quiere decir que en la clasificación obtenida apenas se encuentran similitudes con la categorización real entre pacientes sanos y que padecen glaucoma. Por lo tanto, se concluye que con el algoritmo de clustering DBSCAN no será posible diagnosticar automáticamente la enfermedad con una probabilidad de acierto suficientemente alta. Este hecho, revertirá en la búsqueda de otras técnicas con las que analizar el conjunto de datos para conseguir resultados más satisfactorios.

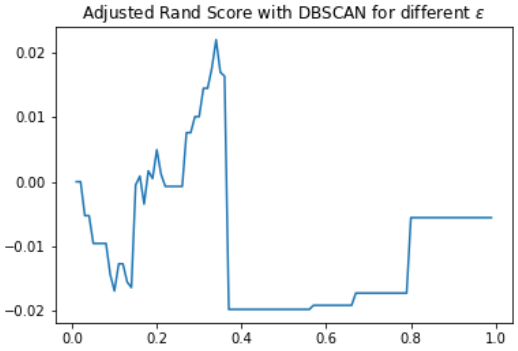


Figura 4.11: ARI 2-distancia (DBSCAN)

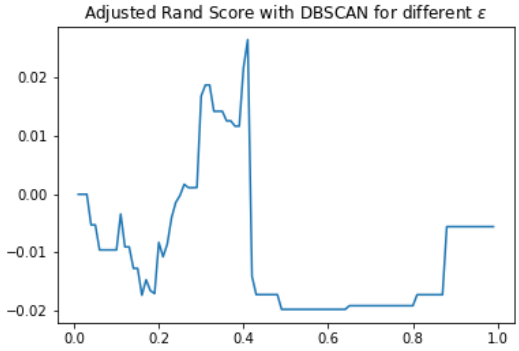


Figura 4.12: ARI manhattan (DBSCAN)

## 4.2. Análisis de componentes principales

Ante la dificultad de diagnosticar el glaucoma utilizando las técnicas de clustering convencionales, resulta interesante estudiar la importancia de cada uno de los momentos geométricos que conforman nuestro sistema. De esta manera es posible conocer la relevancia que presenta cada momento y entender qué papel juega en la descripción de la retina. Así, se consigue una mayor explicabilidad del sistema, lo que revierte en un entendimiento que permite trabajar mejor con los datos que se tienen. Para lograrlo se utiliza el análisis de componentes principales (Smith, 2002), cuyas siglas (PCA) se corresponden con su nomenclatura inglesa *Principal Component Analysis*. En primer lugar, se introduce brevemente esta técnica.

### 4.2.1. Introducción al PCA

Sea un sistema $S \in \mathbb{R}^n$ con $N$ elementos que se denota por $\{(X_i^k)_{i=1}^n\}_{k=1}^N$, así cuando se utiliza $X_i$ se refiere al vector de $N$ componentes formado por la i-ésima variable de los $N$ elementos, y cuando se usa $X^k$ se refiere al k-ésimo elemento con $n$ componentes. De esta forma, entendiendo cada una de las $n$ dimensiones como variables aleatorias, se cuenta con una muestra de tamaño $N$, de tal manera que se puede calcular la covarianza muestral entre ellas utilizando la fórmula que define el estimador insesgado (Gomez Villegas, 2005) de la covarianza de dos variables aleatorias:

$$cov(X_i, X_j) = \frac{\sum_{k=1}^N (X_i^k - \overline{X}_i)(X_j^k - \overline{X}_j)}{N}$$

donde $\overline{X}_i = \frac{1}{N}\sum_{k=1}^N X_i^k$, es decir, $\overline{X}_i$ se corresponde con la media muestral de la variable $X_i$. Si se denota por $\hat{X}_i$ a la variable centrada de $X_i$ respecto a su media muestral $\overline{X}_i$, o lo que es lo mismo, $\hat{X}_i := (\hat{X}_i^k)_{k=1}^N := (X_i^k - \overline{X}_i)_{k=1}^N$, la definición de la covarianza muestral es equivalente a escribir $cov(X_i, X_j) = \frac{1}{N}\langle \hat{X}_i, \hat{X}_j \rangle$ (siendo $\langle v, w \rangle$ el producto escalar usual de $v$ y $w$). Como la varianza muestral (Upton y Cook, 2014), también llamada momento muestral respecto al centro de orden 2, se define como $var(X_i) = \frac{1}{N}\sum_{k=1}^N (X_i^k - \overline{X}_i)^2$ se deduce:

$$var(X_i) = cov(X_i, X_i) = \frac{1}{N}\langle \hat{X}_i, \hat{X}_i \rangle = \frac{1}{N}\|\hat{X}_i\|^2$$

Con estos conceptos es posible definir la matriz cuadrada de orden $n$ llamada matriz de covarianza muestral: $C = (cov(X_i, X_j))_{ij} \in \mathfrak{M}_n(\mathbb{R})$. Asimismo se puede determinar la matriz de correlaciones $R := r_{ij} \in \mathfrak{M}_n(\mathbb{R})$ con:

$$r_{ij} = \frac{cov(X_i, X_j)}{\sqrt{var(X_i)var(X_j)}} = \frac{\langle \hat{X}_i, \hat{X}_j \rangle}{\|\hat{X}_i\| \cdot \|\hat{X}_j\|}$$

Por lo tanto, desde un punto de vista geométrico, se tiene que $r_{ij} = \cos\alpha$ con $\alpha$ el ángulo formado por $\hat{X}_i$ y $\hat{X}_j$.

Con ambas matrices se puede definir un análisis de componentes principales. Utilizando $C$ el método se denomina PCA basado en las covarianzas y usando $R$ se llama PCA basado en las correlaciones. A continuación se detallará este último, debido a que el rango y escala de las variables de nuestro sistema difieren ampliamente (como se ha visto en la sección 3.3).

Como $R$ es una matriz simétrica es diagonalizable y, dado que la suma de los valores propios es equivalente a la traza de la matriz, se tiene que $\sum_{i=1}^n \lambda_i = n$. Por esta razón, estos valores propios también se conocen como pesos de las componentes.

Las componentes principales son los vectores propios de dicha matriz de correlación (que pueden representarse como combinación lineal de las variables del sistema), las cuales es posible ordenar de mayor a menor peso y quedarse con las $p < n$ primeras, reduciendo así la dimensionalidad del sistema.

Un concepto importante con el que se estudian las componentes principales es el de varianza explicada. Una vez se cuenta con cada una de las componentes, es posible hallar las diferentes varianzas de cada una de ellas. Estas están ordenadas de mayor a menor coincidiendo con la ordenación de los valores propios de la matriz, de manera que la primera componente principal contará con la mayor varianza. Si se normalizan dichas varianzas con la suma total de todas ellas y se halla la fracción correspondiente se calcula la varianza explicada de la componente. Como es lógico, si tomamos tantas componentes como variables posee el sistema, la suma de las varianzas explicadas será uno.

### 4.2.2. Resultados del PCA restringido a la excavación

Con el objetivo de entender mejor la composición del sistema de datos e incluso simplificarlo, se procede a la ejecución de un análisis de componentes principales. De esta forma, se obtiene la combinación lineal de las distintas variables (que son los siete momentos geométricos) que generan cada una de las componentes y es posible determinar "la relevancia" (tomando el valor absoluto del coeficiente correspondiente) de cada dimensión sobre el sistema.

Comenzando con el análisis de componentes principales de los momentos geométricos restringidos a la excavación de la ILM, se observa una clara preponderancia de la primera componente principal sobre las demás, ya que esta posee un ratio de varianza explicada del 99,99 %. Por lo tanto la dirección de esta componente, describe, en gran medida, la distribución de los datos de nuestro sistema. De hecho, si se observa la figura 4.13, se deduce que esta dirección se corresponde prácticamente con el primer momento geométrico, es decir, que la primera componente principal $\omega_0$ puede descomponerse de la siguiente forma:

$$\omega_0 = \mu_0 M_0 + \mu_2 M2$$

tal que $M_0$ y $M_2$ se corresponden con el primer y tercer momento geométrico, respectivamente, y se tiene que $|\mu_0|/\sum_{i=0}^{6}|\mu_i| = 0{,}99$ y $|\mu_2|/\sum_{i=0}^{6}|\mu_i| = 0{,}01$.

La interpretación de las equivalencias anteriores es que la dirección de $\omega_0$ coincide prácticamente con la de el primer momento geométrico, con una ínfima dependencia del tercero y sin desviación hacia las direcciones del resto de momentos. Esto quiere decir que la primera componente principal se encuentra en el plano generado por el primer y tercer momento geométrico y prácticamente coincide con el primero. Además, dada el alto ratio de varianza explicada, estas características son generalizables a prácticamente todos los elementos de nuestro sistema de datos. Por lo tanto, si restringimos los momentos geométricos a la zona de excavación de la ILM, una amplia mayoría de elementos se encontrarán alrededor de la recta que tiene la dirección del primer momento geométrico.

Este resultado concuerda con el estudio preliminar de los datos llevado a cabo, ya que, si se fija la atención en la fila (o columna) del primer momento geométrico de la matriz de dispersión de estos momentos restringidos a la excavación (véase la figura 3.6), se puede observar que todos los puntos en función del resto de momentos se encuentran alrededor de una misma vertical (u horizontal si se observa la columna), excepto cuando se empareja el primer momento con el tercero, en cuyo caso todos los elementos siguen encontrándose alrededor de una trayectoria recta ligeramente inclinada. De esta manera, a pesar de no haber vislumbrado una clara separación de los conjuntos de pacientes que

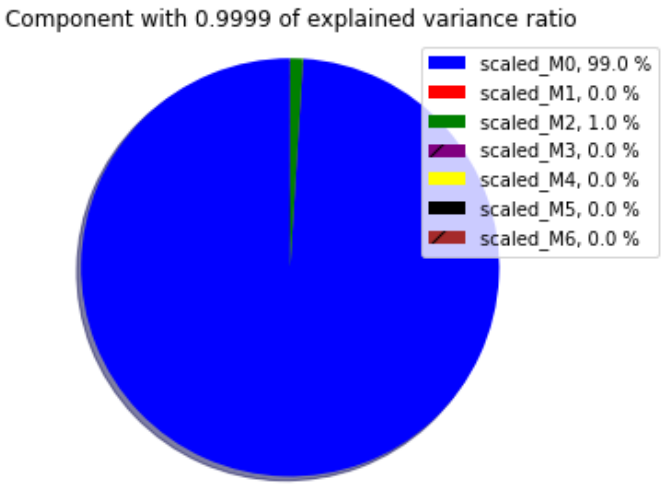


Figura 4.13: Primera componente principal restringiendo los momentos a la excavación

padecen glaucoma y del grupo de control, gracias la matriz de dispersión ha sido posible verificar los resultados de este análisis de componentes principales, así como entender mejor la distribución de los datos de nuestro sistema.

Dado el alto valor del ratio de varianza explicada de esta primera componente y, en consecuencia, el ínfimo valor del ratio de varianza explicada del resto de componentes principales (el ratio de la varianza explicada de la segunda es de 0,01 %), la relevancia del resto de componentes principales afectarán en discreta medida a la distribución general del sistema y las conclusiones que puedan extraerse de ellas tendrán una trascendencia poco significativa. Por estos motivos no se ve necesario llevar a cabo el estudio de estas.

### 4.2.3. Resultados del PCA de toda la ILM

Una vez se ha llevado a cabo un análisis de componentes principales de los momentos geométricos restringidos a la zona de excavación de la ILM y se han obtenido conclusiones que han ayudado a mejorar la comprensión global de nuestro sistema de datos, se presenta el interés de repetir el proceso para estudiar los momentos geométricos describiendo toda la ILM. Por ello, en esta sección se analizan los resultados obtenidos al ejecutar un PCA sobre los momentos geométricos sin restringirlos.

Tras ejecutar el análisis de componentes principales, haciendo uso de la clase *PCA* importada de la librería de Python *sklearn*, de nuevo se vuelve a obtener una primera componente principal con un ratio de varianza explicada muy significativo (en esta ocasión el ratio resultante es del 99,36 %). Este hecho indica que es posible reducir el sistema a una única dimensión (la cual se correspondería con la primera componente principal) sin perder apenas información. De hecho, como ocurre en el caso de restricción de los momentos geométricos, la segunda componente principal posee un ratio de varianza explicada ínfimo (de aproximadamente 0,64 %) y las subsiguientes no presentan una relevancia trascendental.

En cuanto a la combinación lineal que define a esta primera componente principal, en la figura 4.14, se observa que la dirección preponderante vuelve a ser la del primer momento geométrico, aunque ligeramente en menor medida. En este caso tendremos que la primera componente puede descomponerse de la siguiente forma:

$$\omega_0 = \mu_0 M_0 + \mu_1 M_1 + \mu_2 M2$$

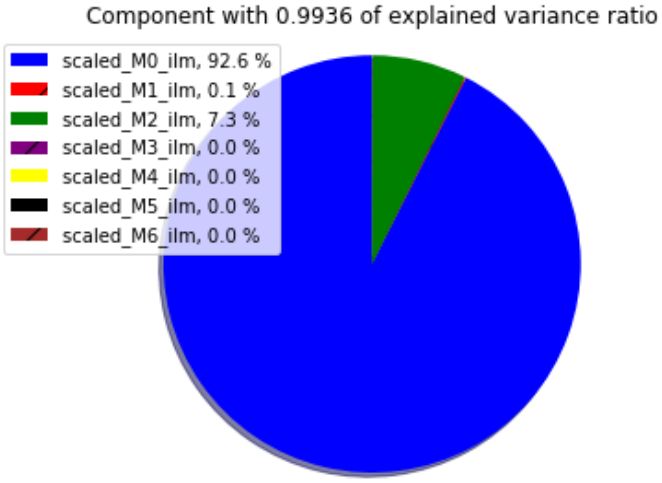


Figura 4.14: Primera componente principal sin restringir los momentos geométricos

tal que se cumple que $|\mu_0|/\sum_{i=0}^{6}|\mu_i| = 0{,}926$, $|\mu_1|/\sum_{i=0}^{6}|\mu_i| = 0{,}001$ y, por último, $|\mu_2|/\sum_{i=0}^{6}|\mu_i| = 0{,}073$.

En este caso, el tener un, ligeramente, menor ratio de varianza explicada y la aparición de una combinación lineal de tres momentos con una mayor relevancia de los dos minoritarios, son factores que complican la extracción de conclusiones tan visuales observando directamente la matriz de dispersión correspondiente (véase la figura 3.7). Este hecho, probablemente se vea generado por un mayor número de elementos en el sistema, lo cual, por otro lado, agudiza visualmente la correlación entre ciertas parejas de momentos geométricos.

Lo que sí es posible afirmar es que la primera componente principal está prácticamente embebida en el plano generado por el primer y tercer momento geométrico (con más relevancia de la dirección del primero). Se encontrará en tres dimensiones distintas debida a una leve influencia del segundo momento geométrico que eleva a $\omega_0$ de dicho plano.

Aunque no resulta significativa la segunda componente principal (por su bajo ratio de varianza explicada) como para estudiarla en profundidad, resulta curioso señalar que en ella se "invierten los papeles que juegan" el primer y tercer momento geométrico, es decir, el valor absoluto del coeficiente lineal del tercero dividido por la suma normalizada de los valores absolutos de los coeficientes lineales representa aproximadamente el 91,1 %, mientras que si colocamos el valor absoluto del coeficiente lineal del primer momento geométrico en el numerador obtenemos aproximadamente el 7,2 %. Por otro lado, el valor ponderado de importancia del segundo momento aumenta levemente a aproximadamente el 1,4 %.

### 4.2.4. Conclusiones del PCA

Tras haber llevado a cabo el análisis de componentes principales tanto sobre los momentos geométricos restringidos a la excavación como describiendo toda la ILM, se pueden señalar algunas observaciones con las que se encuentran posibles hipótesis como trabajo futuro y con las que se continúa en este estudio.

En primer lugar, se ha observado que la primera componente principal poseía un ratio de varianza explicada preponderante en ambos casos. Sin embargo, cuando los momentos eran

aplicados a toda la ILM comenzaba a aumentar, levemente, el ratio de varianza explicada de la segunda componente (mientras se reducía el de la primera). Este comportamiento podría ser simplemente un caso aislado y no verse representado con datos distintos (de manera que se conserve la preponderancia de la primera componente princial), o tener su justificación en el estudio de la ILM completa, el aumento de tamaño del conjunto de datos o ambas circunstancias. Para abordar esta cuestión sería necesario contar con un sistema con mayor número de elementos.

Otra observación que podría suscitar interés, y con ciertas similitudes a la anterior, resulta del aumento de importancia ponderada del segundo y tercer momento geométrico en la primera componente principal en el caso en el que se analiza toda la ILM. Cuanto más disperso sea un sistema de datos, por lo general, más distribuida se encuentra la influencia de las variables en las primeras componentes. Sin embargo, no es posible afirmar con absoluta certeza que este aumento de relevancia se deba exclusivamente al mayor número de elementos que posee el sistema, ya que el hecho de no restringir los momentos geométricos podría tener trascendencia en esta cuestión. Sin embargo, al igual que ocurre con la observación anterior, para el estudio de este tema debe contarse con un conjunto de datos de mayor tamaño.

Por último, cabe destacar que, a pesar de haber sido útil para mejorar la comprensión sobre cómo están distribuidos los elementos del sistema, el análisis de componentes principales tiene en cuenta todos los datos independientemente del conjunto al que pertenezcan (controles o pacientes de glaucoma). Por esta razón, no nos permite dirimir la relevancia que tiene cada momento geométrico a la hora de clasificar los elementos del sistema entre pacientes sanos y que padecen la enfermedad. Este problema, obliga a la búsqueda de técnicas de aprendizaje automático que proporcionen dicha información. En este estudio se utiliza la llamada *importancia de Gini* en los *Árboles de decisión*.

## 4.3. Árboles de decisión

A diferencia del análisis de componentes principales, los árboles de decisión son un método de aprendizaje automático que tiene en cuenta la clasificación dada del conjunto de datos. A pesar de que, generalmente, no son utilizados con el objetivo de reducir las dimensiones del sistema o ponderar la importancia de cada una de ellas, algunos investigadores como Sugumaran et al. (2007) y Cho y Kurup (2011) han hecho uso de ellos para seleccionar las variables con mayor potencial con el fin de clasificar los distintos elementos. Un árbol de decisión (Rokach y Maimon, 2008) es un modelo de predicción que, dado un conjunto de datos, construye un diagrama de reglas lógicas, muy similar a los sistemas de predicción basados en reglas. Estos diagramas se utilizan para clasificar los distintos elementos del conjunto de predicción en función de una serie de características con las que se ha entrenado previamente el modelo a través de un conjunto de datos de entrenamiento. Existen diversos algoritmos que implementan los árboles de decisión. En este estudio, utilizaremos los recursos facilitados por la librería de Python *sklearn*, cuya implementación es una versión optimizada del algoritmo CART (Breiman et al., 1984) con la entropía como criterio de bifurcación.

Entre las ventajas de los árboles de decisión se encuentra el hecho de que tienen en cuenta la categorización dada del conjunto de datos, ya que se trata de un método de clasificación de aprendizaje automático supervisado, y que son una de las técnicas más explicables de machine learning. Por estos motivos, los árboles de decisión son una técnica que permite determinar cuál de los momentos distingue mejor entre los pacientes del grupo de control y que padecen glaucoma y con los que es posible presentar un modelo de

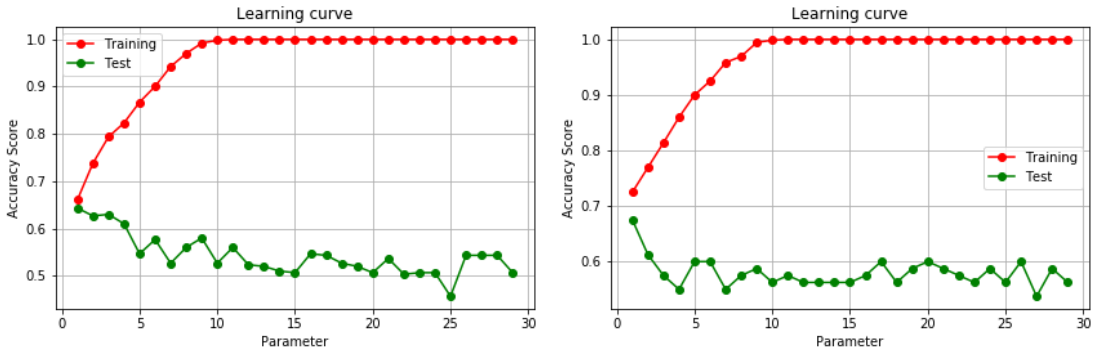


Figura 4.15: Ejemplos de curva de aprendizaje

diagnóstico y estimar su eficacia.

### 4.3.1. Determinar la profundidad óptima

Uno de los parámetros de inicialización de un árbol de decisión es la profundidad máxima que debe tener. Conforme aumenta, el modelo de entrenamiento tiene en cuenta un mayor número de casos. No obstante, si se inicializa con un valor demasiado alto se puede producir un sobreaprendizaje, lo cual empeora el comportamiento y la eficacia del árbol de decisión. Por esta razón, es importante calcular una profundidad óptima para el árbol de decisión. Para llevar a cabo esta tarea, se ejecuta numerosas veces el modelo de entrenamiento con validación cruzada buscando en cada una de ellas la profundidad que obtiene el mejor resultado en la curva de aprendizaje (sin alcanzar el sobreaprendizaje) y la mayor precisión con el conjunto de prueba.

Por ejemplo, en la figura 4.15, se pueden observar dos curvas de aprendizaje distintas (en rojo) con sus respectivas gráficas que relacionan la profundidad con la precisión obtenida al evaluar el conjunto de entrenamiento (en verde). Con el fin de evitar el sobreaprendizaje (que se produce cuando el valor de precisión del conjunto de entrenamiento es 1) y la pérdida de información (se considera que no se tienen en cuenta todas las casuísticas más representativas cuando el valor de precisión del conjunto de entrenamiento es inferior a 0.9), se escoge una profundidad cuyo valor de precisión se encuentre en el intervalo $(0{,}9, 1)$. Como todas estos posibles valores del parámetro inicializador son razonables en lo que al conjunto de entrenamiento se refiere, entre ellos se tomará el que alcance el valor máximo

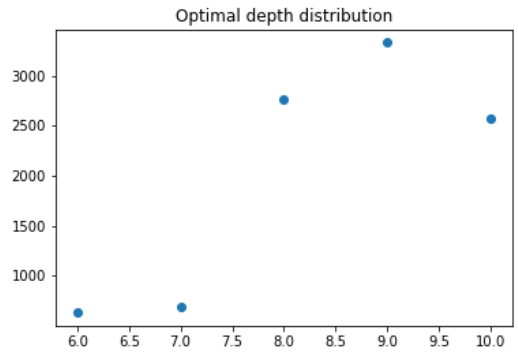


Figura 4.16: Distribución (ILM)

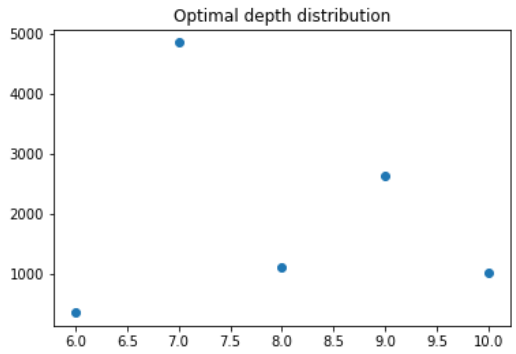


Figura 4.17: Distribución (excavación)

de precisión en el conjunto de prueba (o conjunto de *test*). De esta manera en la gráfica de la izquierda de la figura 4.15, la profundidad óptima sería 9 (ante la misma precisión en el conjunto de test se priorizará la que obtenga mayor precisión en el conjunto de entrenamiento), mientras que en la imagen de la derecha se correspondería con 6.

Con este criterio de elección de la profundidad óptima, y dado que utilizando la entropía un árbol de decisión varía su resultado por la pseudo-aleatoriedad intrínseca en su algoritmo de generación (Breiman et al., 1984), se generan 10000 curvas de aprendizaje (como las que pueden observarse en la figura 4.15) para de ellas extraer la distribución de dicho parámetro de inicialización óptimo. Este estudio se lleva a cabo tanto para el conjunto de datos descrito por los momentos geométricos aplicados a toda la ILM (el resultado se muestra en la figura 4.16), como al sistema en el que los invariantes se restringen a la excavación (la distribución se presenta en la figura 4.17). Observando tanto la representación gráfica de las distribuciones de la profundidad óptima como sus estadísticos muestrales descriptivos (que se presentan en la tabla 4.1, en la cual *std* hace referencia a las siglas en inglés de la desviación típica) de ambos casos, se elegirán las profundidades de 9 y 7 para generar los árboles de decisión cuando los momentos geométricos describen toda la ILM y cuando los invariantes se restringen a la excavación, respectivamente.

| | **media** | **std** | **mín** | **máx** | **moda** | **25 %** | **50 %** | **75 %** |
|---|---|---|---|---|---|---|---|---|
| **ILM** | 8.651 | 1.124 | 6 | 10 | 9 | 8 | 9 | 10 |
| **Excavación** | 7.9098 | 1.1401 | 6 | 10 | 7 | 7 | 7 | 9 |

Tabla 4.1: Estadísticos descriptivos de las distribuciones de la profundidad óptima

### 4.3.2. Importancia de Gini

Una vez se han obtenido las profundidades óptimas de los árboles de decisión sobre ambos conjuntos, es posible comenzar con el estudio de los datos con esta técnica de aprendizaje automático. Una de las cuestiones que puede suscitar interés es determinar qué momentos son más "importantes" (concepto que se formaliza más adelante) a la hora de diagnosticar el glaucoma, es decir, cuando queremos distinguir entre individuos pertenecientes al grupo de control o pacientes glaucomatosos, qué momentos aportan mayor información para prestarles mayor atención. En los árboles de decisión, cada nodo viene determinado por un invariante, de manera que la pregunta pasa a ser cómo cuantificar la importancia de un nodo.

Una buena intuición puede ser pensar que, para estudiar la relevancia de un nodo, es importante tener en cuenta su profundidad en el árbol, ya que cuanto menor sea esta mayor es el número de elementos que diferencia. No obstante, no sería muy representativo si solo dividiera elementos de la misma clase o no redujera la entropía de la muestra. Por lo tanto, el número de individuos que alcanzan el nodo y la categoría de estos también son factores a considerar. Sin embargo, tampoco representan elementos determinantes si se mantiene la proporción de cada categoría en los nodos hijo. Si se continua con esta reflexión, es posible revelar una notable cantidad de parámetros que pueden tener relevancia en la definición de importancia de un nodo en un árbol de decisión. En este estudio, se usa la *importancia de Gini* (Breiman, 2001), la cual se define formalmente con la siguiente expresión:

$$I(n) = \omega_n H_n - \omega_{left(n)} H_{left(n)} - \omega_{right(n)} H_{right(n)}$$

donde $I(n)$ es la importancia de Gini (cuyas siglas son GI por su nombre en inglés) del nodo $n$, $\omega_n$ es el número ponderado de muestras que alcanzan el nodo $n$, $H_n$ es la entropía

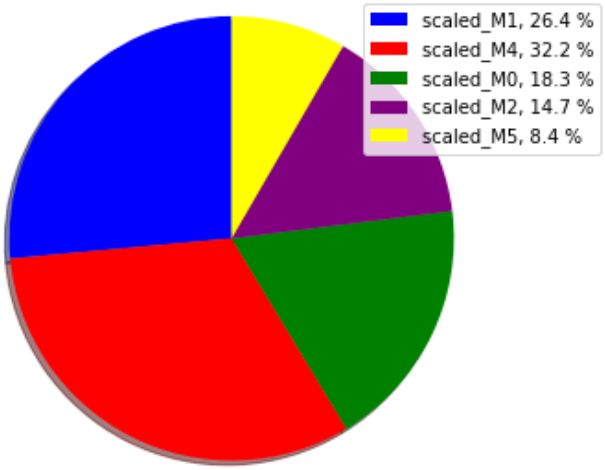


Figura 4.18: Distribución de la importancia de Gini del árbol de la figura 4.19

del nodo $n$, $left(n)$ es el hijo izquierdo del nodo $n$ y $right(n)$ es el hijo derecho del nodo $n$. En consecuencia, la importancia de cada variable (en este caso de cada momento) se define con la siguiente fórmula:

$$I(M_i) = \frac{\sum_{n \in Nod(M_i)} I(n)}{\sum_{n \in Nod} I(n)}$$

donde $I(M_i)$ es la importancia de Gini del momento $M_i$, $Nod(M_i)$ es el conjunto de nodos que se dividen en función de $M_i$ y $Nod$ es el conjunto de todos los nodos.

En este trabajo se utiliza la importancia de Gini del momento $M_i$ normalizada, es decir, el valor $I(M_i)/\sum_{j=0}^{6} I(M_j)$. Así dado el árbol de ejemplo de la figura 4.19, se obtiene la distribución de la importancia de Gini de los momentos representada en la figura 4.18. Si se observa este ejemplo, se verá que el momento $M_6$ no es determinante en la clasificación de los elementos, por lo que su importancia de Gini es nula. Además se ve que en la raíz del árbol el momento $M_1$ casi es capaz de separar diecisiete pacientes glaucomatosos del conjunto, por lo que su importancia de Gini es muy alta. Por el contrario, $M_5$ solo aparece en el nivel más profundo del árbol y no consigue reducir notablemente la entropía del conjunto de muestras, lo que revierte en una importancia de Gini de este invariante muy baja.

Tras haber comprendido el funcionamiento de la importancia de Gini, gracias al ejemplo práctico anterior, se utiliza para cuantificar la relevancia de los distintos momentos. Con este fin, se implementa el cálculo de este valor y se generan un millón de árboles de decisión calculando la distribución de la importancia de Gini de cada uno de ellos. Esta acción se lleva a cabo tanto sobre el conjunto de elementos descritos por los momentos geométricos aplicados a toda la ILM, como el sistema constituido por los invariantes restringidos a la excavación. De esta manera se pueden obtener distribuciones de las importancias de cada uno de los momentos en cada uno de los casos.

Para los momentos geométricos restringidos a la excavación se obtiene la distribución que se representa en la figura 4.20. Como puede observarse, en todos los árboles de decisión generados el cuarto ($M_3$) y séptimo ($M_6$) momento presentan una importancia nula. Por lo tanto, podemos afirmar con certeza que, con los datos dados, se podría prescindir de ellos para diagnosticar el glaucoma cuando se restringen los invariantes a la cúpula. Esto es

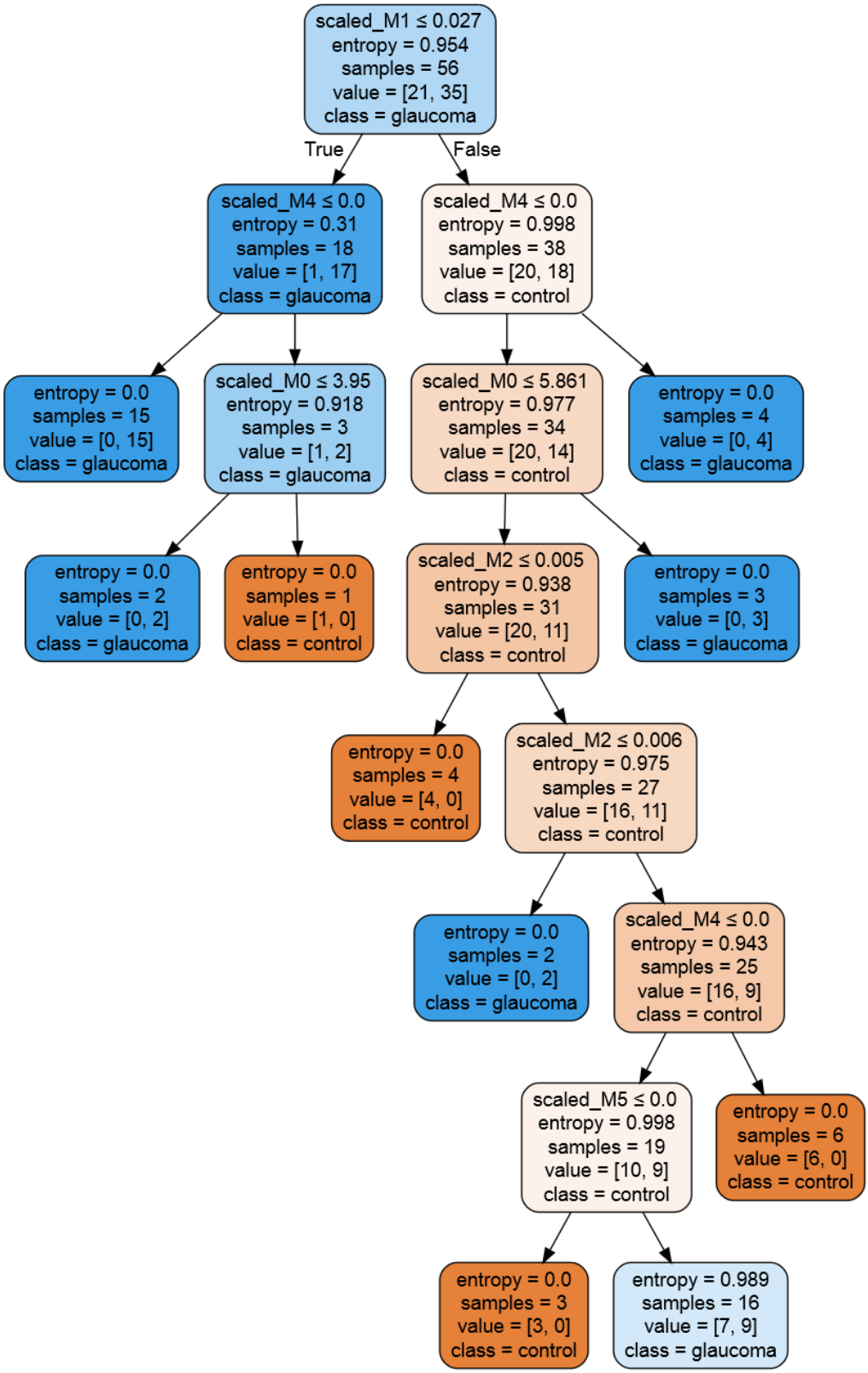


Figura 4.19: Árbol de decisión que genera la importancia de Gini distribuida según la figura 4.18

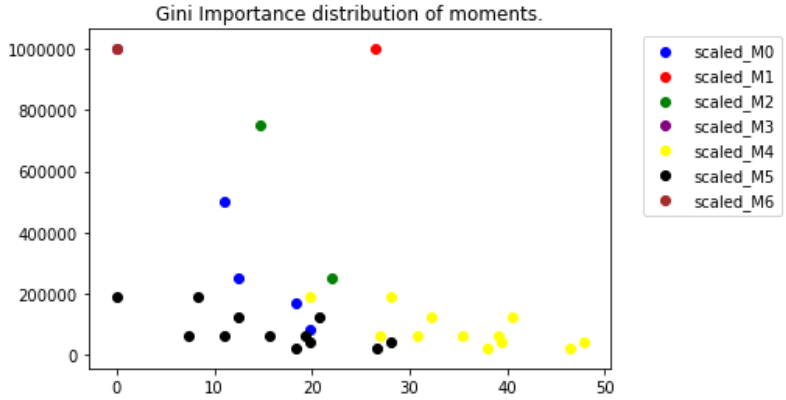


Figura 4.20: Distribución de la importancia de Gini de cada momento restringido a la excavación

consistente con lo deducido en la sección 3.4.1, en la que, gracias a la figura 3.4, se observaba una correlación lineal entre los momentos $M_1$, $M_3$ y $M_6$. Como el invariante $M_1$ presenta una importancia considerable y constante en todos los árboles (el valor aproximado de la importancia de Gini es de 26,449) y $M_3$ y $M_6$ poseen una correlación lineal con $M_1$, es posible prescindir de ellos.

Por otro lado es posible observar momentos con una desviación típica bastante significativa como $M_4$ y $M_5$ (que poseen desviaciones típicas con valor 7.98). Como se puede deducir a partir de la figura, el primero es el invariante más relevante en la tarea de diagnóstico (con una relevancia media de 31.71 %), mientras que el segundo apenas es significativo (su media aproximada tiene un valor del 12 % de GI). Esto vuelve a estar en consonancia con lo visto en la sección 3.4.1, ya que, como estos invariantes están fuertemente correlacionados, es sensato observar que la alta GI de uno vaya en detrimento de la relevancia del otro. Por último, cabe señalar que la importancia media de Gini de los momentos $M_2$ y $M_0$ (con valores aproximados de 16.5 % y 13.3 %) los sitúan en el tercer y cuarto invariantes más significativos a la hora de clasificar los elementos del sistema.

Sorprendentemente, las conclusiones varían radicalmente cuando se lleva a cabo el estudio de los momentos geométricos aplicados a toda la ILM. Como puede observarse en la figura 4.21, el séptimo invariante $M_6$ es el que posee la mayor importancia de Gini (con un valor medio de 24 %). En consecuencia, y de nuevo en consonancia con las correlaciones

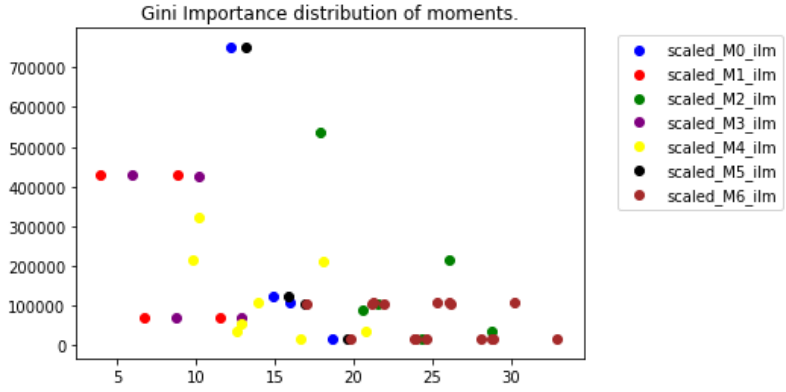


Figura 4.21: Distribución de la importancia de Gini de cada momento sobre toda la ILM

observadas en la sección 3.4.1, los momentos correlacionados con $M_6$ (que son $M_1$ y $M_3$) presentan los valores de GI más bajos (sus medias son 6.78 % y 8.45 %, respectivamente, aunque ninguno presenta una GI nula en ningún árbol generado). Por el contrario, y de manera similar al caso anterior, en esta ocasión $M_2$ también resulta tener una relevancia significativa en comparación con el resto de invariantes (ocupa el segundo lugar en orden de GI media con una importancia media de 20.76 %). Ambos casos también son bastante similares en la GI del primer y sexto momento geométrico, ya que en este caso poseen una importancia media de 13 % y 14 %, respectivamente, con ambos una desviación típica de 1.56. No obstante, difieren en que la desviación típica es mucho menor en la distribución de la figura 4.21, ya que su valor máximo es de 3.63 (frente a 7.98 restringiendo los momentos), y que el quinto invariante muestra una GI media mucho menor (en este caso es de 12.92 % frente al 31.71 % anterior).

Por lo tanto, como se ha podido observar, la importancia de los momentos varía notablemente en función de si describen toda la ILM o se restringen a la excavación. Sin embargo, en ambos casos las correlaciones entre los invariantes genera que sea posible mantener uno y prescindir del resto de momentos correlacionados para la automatización de la tarea de diagnóstico.

### 4.3.3. Modelo de predicción

El objetivo principal de los árboles de decisión no es la cuantificación de la importancia de las distintas variables del sistema, sino que son modelos de predicción que permiten clasificar los elementos del conjunto de datos entrenándolos previamente. Por lo tanto, es posible generar estas estructuras para tratar de diagnosticar la enfermedad. Para ello, se dividirá el sistema en conjunto de datos de entrenamiento, que representará un 70 % del total, y de prueba, el 30 % restante. A continuación, se presentan los resultados (utilizando las profundidades óptimas calculadas en la sección 4.3.1) obtenidos tanto restringiendo los momentos geométricos a la excavación como aplicándolos sobre toda la ILM.

El árbol de decisión entrenado que se genera cuando se utiliza el conjunto de datos descrito por los momentos geométricos restringidos a la excavación se muestra en la figura 4.24. Como puede observarse, cumple con la distribución de importancia de invariantes presentada en la sección anterior, ya que el momento $M_1$ presenta bastante importancia mientras que $M_3$ y $M_6$ (que están correlacionados con el segundo invariante) poseen una relevancia nula. Asimismo, $M_5$ resulta muy representativo, mientras que $M_4$ (que está correlacionado con $M_5$) queda relegado a un nodo en la máxima profundidad. Por otro lado, $M_2$ y $M_0$ mantienen una importancia notable como se observaba en la figura 4.20.

Si se evalúa este modelo sobre el conjunto de datos de prueba se obtiene la matriz de confusión (que nos indica tanto los aciertos en la predicción como los falsos negativos y positivos) de la figura 4.22. En ella se observa un alto ratio de confianza para diagnosticar pacientes glaucomatosos (con una eficacia del 75 %), sin embargo una precisión muy baja para detectar los individuos pertenecientes al grupo de control (una eficacia del 40 %, lo cual posee un error que no es asumible para un programa de diagnóstico). Este resultado se origina por una clara influencia por el mayor número de pacientes de glaucoma (35 frente a 21) que condiciona el árbol de decisión. Esto indica que, debido al pequeño número de datos y un desequilibrio suficientemente notable en ellos, no es posible usar este modelo para la automatización del diagnóstico. Sin embargo, no se descarta esta técnicas de aprendizaje automático en unas condiciones en las que se posean un sistema más adecuado para ella.

En el caso de los momentos geométricos aplicados a toda la ILM, se genera el árbol de decisión que se muestra en la figura 4.25. Al igual que ocurría restringiendo los invariantes

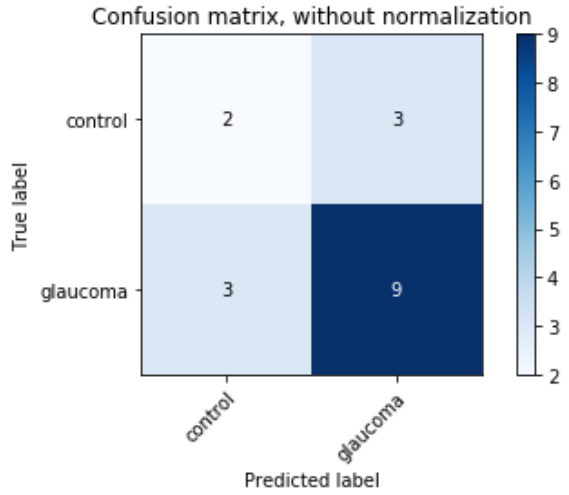


Figura 4.22: Matriz de confusión del árbol de la figura 4.24

a la excavación, se mantiene la distribución de importancias descrita en la sección 4.3.2, lo cual reafirma las conclusiones presentadas en ella. Sin embargo, los resultados de su matriz de confusión (véase la figura 4.23) presenta una significativa mejoría respecto a los del caso anterior. En esta ocasión, los aciertos de predicción y la eficacia entre pacientes glaucomatosos y miembros del grupo de control está más equilibrada (las eficacias son 62 % y 55 %, respectivamente), ya que, aunque el número de elementos en un conjunto y en otro continúa presentando cierto desequilibrio (35 individuos en el grupo de control frente a 45 personas que padecen glaucoma), proporcionalmente son más similares. Sin embargo, el ratio de confianza continúa siendo excesivamente bajo para una aplicación de diagnóstico. Uno de los factores que podría ser muy influyente es la pequeña cantidad de elementos en el sistema, que podría producir que algunas casuísticas no estuvieran contempladas o se encontraran infrarrepresentadas.

A pesar de no haber podido obtener modelos de predicción eficaces, los árboles de decisión han permitido comprender mejor el sistema de datos con el que se cuenta mediante el estudio de la importancia de Gini de los distintos momentos.

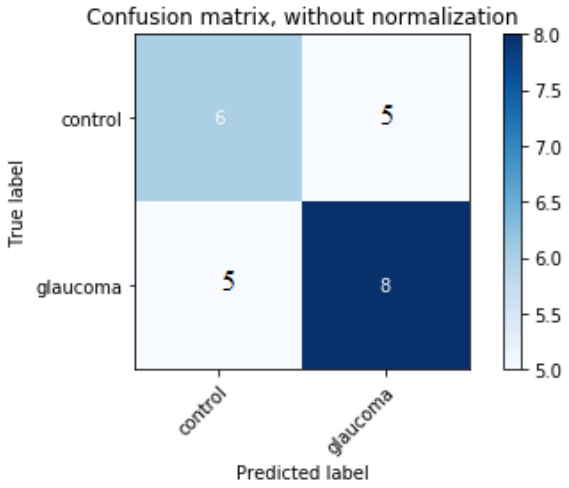


Figura 4.23: Matriz de confusión del árbol de la figura 4.25


Figura 4.24: Árbol de decisión para los momentos restringidos a la excavación

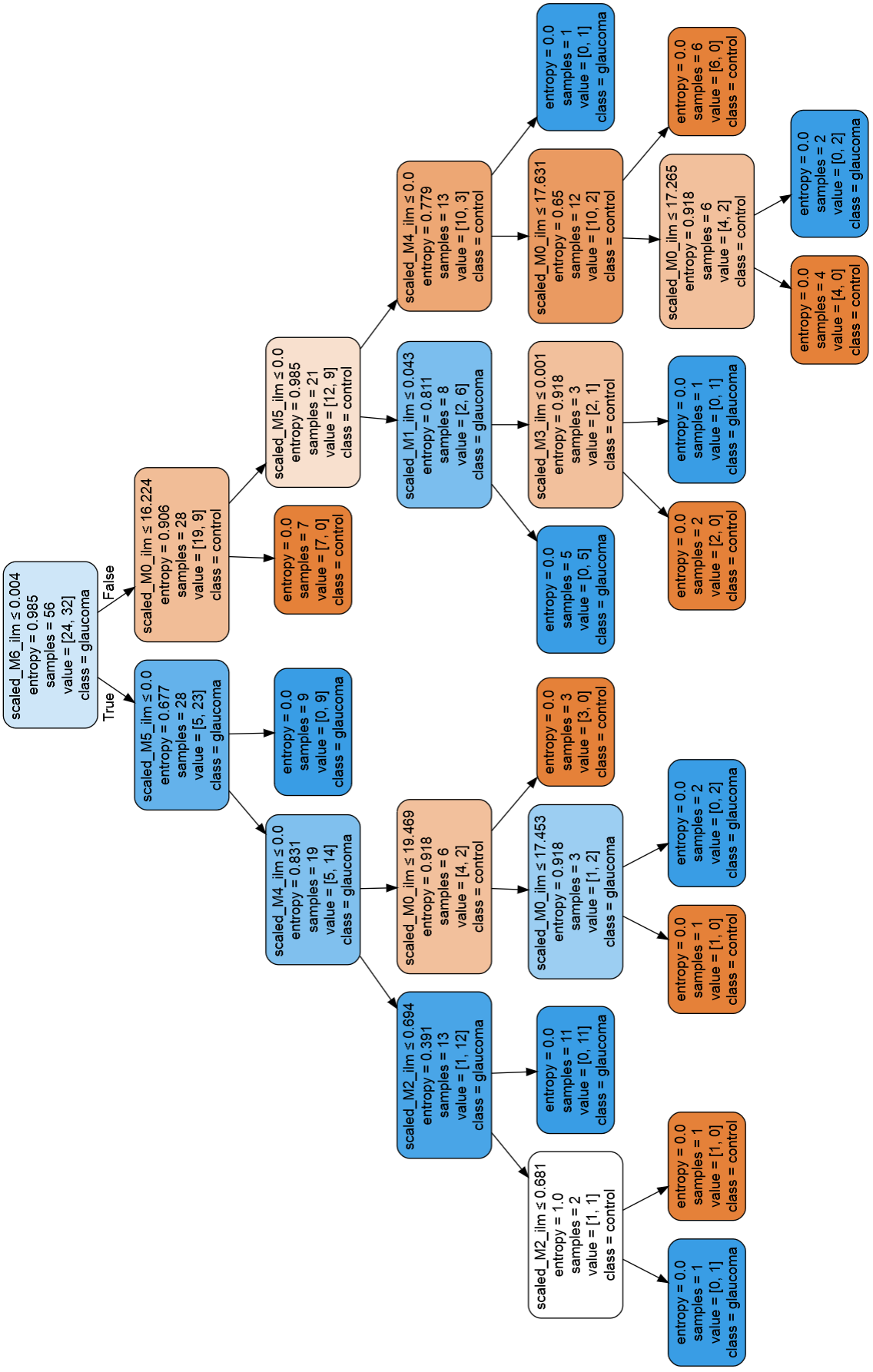


Figura 4.25: Árbol de decisión para los momentos aplicados sobre toda la ILM

Capítulo 5

# Conclusiones y trabajo futuro

*"La vida es el arte de sacar conclusiones suficientes a partir de datos insuficientes."*
— Samuel Butler (1835-1902)

Tras el desarrollo del trabajo, en este capítulo se presentan las conclusiones que es posible extraer de este estudio. Estas se explican en la sección 5.1. Seguidamente, las posibles opciones para la continuación de es te trabajo se presentan en la sección 5.2, con el objetivo de continuar estudiando los momentos geométricos resultantes de las tomografías de coherencia óptica y llegar a conclusiones que aporten información sobre el glaucoma con la cual se pueda desarrollar líneas de investigación en el área de la oftalmología.

## 5.1. Conclusiones

El glaucoma es la primera causa de ceguera irreversible a nivel mundial, a pesar de que la agudeza visual se conserva hasta fases avanzadas de esta enfermedad. Por este motivo, la detección de pacientes glaucomatosos se convierte en un problema sanitario importante. Por lo tanto, el estudio de esta neuropatía óptica resulta muy beneficioso para una considerable parte de la población que la padece y presenta aplicaciones directas en el área de la oftalmología.

El glaucoma se caracteriza por un aumento de la excavación que se encuentra en la papila óptica del ojo humano. Esta depresión puede ser parametrizada a través de momentos geométricos invariantes ante transformaciones de semejanza. En este estudio se han utilizado siete momentos que presentan Xu y Li (2006) en su artículo de investigación. En este trabajo se ha hecho uso de ellos para describir la cúpula del punto ciego de la retina ocular. Estos invariantes han sido analizados tanto con técnicas de estadística básicas como con métodos de aprendizaje automático.

En primer lugar, se ha estudiado la consistencia de los momentos geométricos dado que se contaba con tomografía duplicadas, es decir, pruebas realizadas al mismo paciente en dos ocasiones el mismo día. Para el análisis de esta propiedad se ha cuantificado tanto el error cuadrático medio como la distancia euclídea de los momentos asociados a una de las tomografías duplicadas al resto de elementos del sistema. Sin embargo, se ha demostrado que el conjunto de los siete invariantes no presenta la propiedad de consistencia ni en el caso en el que los momentos se restringen a la excavación ni cuando estos describen toda la ILM. Por lo tanto, a menos que los pacientes hayan sido sometidos a tratamiento médico entre las dos pruebas, es posible sospechar que alguno de ellos no es consistente y podría

estar afectando al resto que lo son. A esto se suma que, como se ha visto más adelante, se ha demostrado la redundancia de ciertos momentos geométricos (dada su correlación) y, además, se cuenta con una muestra de tamaño reducido.

A continuación, se han analizado los estadísticos descriptivos más comunes: media, desviación típica, valor mínimo, valor máximo y cuartiles. Con ello se observó que los momentos geométricos son muy distintos en lo que se refiere a órdenes de magnitud. Por lo tanto, debían escalarse posteriormente para aplicar algunas técnicas de aprendizaje automático.

El último análisis de estadística básica que se ha llevado a cabo es el estudio de correlación y dispersión de los momentos geométricos. Con él se detectaron fuertes correlaciones lineales entre algunos de ellos, lo cual permite concluir que sería posible prescindir de un subconjunto de los invariantes y quedarse con aquellos no correlacionados entre ellos.

Una vez se ha llevado a cabo un estudio preliminar de los datos con técnicas de estadística básica, gracias al cual ha sido posible dilucidar características concretas de los momentos geométricos, se procedió a la utilización de métodos más complejos pertenecientes a la rama de machine learning para abordar cuestiones más complejas. Se comienza con una aproximación utilizando algoritmos de clustering (K-Medias y DBSCAN). Con la utilización de estas técnicas se trata de conocer si el sistema está distribuido de manera que sea posible agrupar los elementos y que estos subconjuntos posean individuos de la misma categoría (pacientes glaucomatosos y pertenecientes al grupo de control). Con ello es posible hacerse una idea de cómo se estructura el conjunto de datos. No obstante, tras ejecutar los algoritmos, se observa que los distintos elementos no se sitúan en un entorno rodeado de individuos de su misma categoría, lo que complica la tarea de automatización de diagnóstico de la enfermedad.

Con el objetivo de conocer la información estructural del sistema de datos, se continúa el trabajo con un análisis de componentes principales. Esta técnica de aprendizaje automático saca a la luz la importancia del primer momento geométrico en el conjunto, ya que determina la dirección sobra la que se sitúan los elementos. Además, gracias al PCA, es posible explicar la representación gráfica mostrada en muchas celdas de la matriz de dispersión generada anteriormente. Asimismo, tras el análisis de componentes principales, es posible describir con mayor precisión cómo se posicionan los elementos del sistema en el espacio.

Por último, se utilizan los árboles de decisión tanto para cuantificar la relevancia que poseen los distintos momentos a la hora de diagnosticar la enfermedad (mediante la importancia de Gini) como para generar un modelo de clasificación de los pacientes. Gracias a ellos, se observa que, dependiendo de si se utilizan los momentos para describir toda la ILM o se restringen los invariantes a la excavación, se debe prestar atención a un subconjunto distinto de ellos para determinar si el paciente padece glaucoma. Además, el comportamiento observado reafirma las correlaciones encontradas anteriormente, por lo que aseguran la base estadística inicial sobre la que se comienza el trabajo. Sin embargo, dado cierto desequilibrio entre las dos categorías del sistema, los modelos de predicción presentan un ratio de confianza demasiado bajo como para generar un diagnóstico automático. Por otro lado, han sido de utilidad para corroborar a las conclusiones de relevancia de los distintos momentos, alcanzadas gracias a la importancia de Gini.

## 5.2. Trabajo futuro

El alcance del trabajo ha estado limitado por el conjunto de datos con el que se contaba. Su pequeño tamaño complicaba la utilización de técnicas estadísticas y de machine learning

para extraer conclusiones con relevancia significativa. Sin embargo, se han observado ciertas características y comportamientos que dan lugar a nuevas cuestiones que pueden abordarse utilizando los resultados presentados en este estudio.

Además de la posibilidad de repetir el estudio llevado a cabo en este trabajo cuando se disponga de una conjunto de datos de tamaño significativo y comprobar así la consistencia de las propiedades y conclusiones presentadas, es posible plantearse otras cuestiones distintas como si se mantendrían los resultados prescindiendo de aquellos momentos que están correlacionados linealmente con otros. Este tema permitiría reducir la dimensionalidad del sistema en caso de que se mantuvieran las propiedades demostradas y no cambiara la estructura en el espacio del conjunto de datos.

Por otro lado, también es inevitable preguntarse si se pudieran obtener nuevas conclusiones distintas y cómo se comportarían los datos al ser sometidos a otras técnicas de aprendizaje automático como máquinas de vectores de soporte o redes neuronales. Asimismo, es posible ampliar el estudio realizado con árboles de decisión utilizando la técnica de *boosting* conocida como *random forest*. Con ella podría obtenerse un modelo de clasificación distinto y que, quizás, obtuviera un ratio de confianza mayor.

Como se ha comentado anteriormente, podría suscitar interés el estudio de la inexistencia de consistencia de los momentos geométricos. Para buscar las razones por las que pueda producirse esto, se podría obtener más información utilizando otras métricas distintas a las presentadas (distancia euclídea y error cuadrático medio), así como eliminar dimensiones del vector de invariantes para investigar si se trata de un subconjunto de ellos los que no mantienen la propiedad de consistencia.

Estas cuestiones planteadas, así como las que se expresaron en la sección 4.2.4, son algunas posibles líneas de investigación por las que podría continuarse trabajando y que estarían correctamente apoyadas por el trabajo presentado.

# Bibliografía

*Y así, del mucho leer y del poco dormir, se le secó el celebro de manera que vino a perder el juicio.*

Miguel de Cervantes Saavedra


Ankerst, M., Breunig, M. M., Kriegel, H.-P. y Sander, J. Optics: ordering points to identify the clustering structure. *ACM Sigmod record*, vol. 28(2), páginas 49–60, 1999.

Benesty, J., Chen, J., Huang, Y. y Cohen, I. Pearson correlation coefficient. En *Noise reduction in speech processing*, páginas 1–4. Springer, 2009.

Breiman, L. Random forests. *Machine learning*, vol. 45(1), páginas 5–32, 2001.

Breiman, L., Friedman, J., Stone, C. J. y Olshen, R. A. *Classification and regression trees*. CRC press, 1984.

Cho, J. H. y Kurup, P. U. Decision tree approach for classification and dimensionality reduction of electronic nose data. *Sensors and Actuators B: Chemical*, vol. 160(1), páginas 542–548, 2011.

Cyganski, D. y Orr, J. A. Applications of tensor theory to object recognition and orientation determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), páginas 662–673, 1985.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. En *Kdd*, vol. 96, páginas 226–231. 1996.

Faber, T. y Stokely, E. Affine transform determination for 3-d objects: A medical imaging application. *Proc. IEEE Comput. Vision Pattern Recognition*, páginas 440–445, 1986.

Gates, A. J. y Ahn, Y.-Y. The impact of random models on clustering similarity. *The Journal of Machine Learning Research*, vol. 18(1), páginas 3049–3076, 2017.

Gomez Villegas, M. A. *Inferencia estadística*. Ediciones Díaz de Santos, 2005.

Grupo, C. Manual cto. *Psiquiatría. 8th ed.: CTO editorial*, 2011.

Hamerly, G. y Elkan, C. Alternatives to the k-means algorithm that find better clusterings. En *Proceedings of the eleventh international conference on Information and knowledge management*, páginas 600–607. 2002.

Hartigan, J. A. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

Hu, M.-K. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, vol. 8(2), páginas 179–187, 1962.

Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., Hee, M. R., Flotte, T., Gregory, K., Puliafito, C. A. et al. Optical coherence tomography. *science*, vol. 254(5035), páginas 1178–1181, 1991.

Kaufman, L. y Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

Li, B. The moment calculation of polyhedra. *Pattern Recognition*, vol. 26(8), páginas 1229–1233, 1993.

Lo, C.-H. y Don, H.-S. 3-d moment forms: their construction and application to object identification and positioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11(10), páginas 1053–1064, 1989.

Mainster, M. A., Timberlake, G. T., Webb, R. H. y Hughes, G. W. Scanning laser ophthalmoscopy: clinical applications. *Ophthalmology*, vol. 89(7), páginas 852–857, 1982.

McKinney, W. et al. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, vol. 14(9), 2011.

Michelessi, M., Lucenteforte, E., Oddone, F., Brazzelli, M., Parravano, M., Franchi, S., Ng, S. M. y Virgili, G. Optic nerve head and fibre layer imaging for diagnosing glaucoma. *Cochrane Database of Systematic Reviews*, (11), 2015.

Netter, F. H. *Atlas de anatomía humana*. Elsevier, 2011.

Noguera, J. M. Oftalmología clínica de jack j. kanski y brad bowling. *Medicina balear*, vol. 27(2), página 51, 2012.

Pavan-Langston, D. *Manual of ocular diagnosis and therapy*. Lippincott Williams & Wilkins, 2008.

Potter, K., Hagen, H., Kerren, A. y Dannenmann, P. Methods for presenting statistical information: The box plot. *Visualization of large and unstructured data sets*, vol. 4, páginas 97–106, 2006.

Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, vol. 66(336), páginas 846–850, 1971.

Reddi, S. Radial and angular moment invariants for image identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2), páginas 240–242, 1981.

Rodríguez, C. G. Técnicas de visión computerizada y reconstrucción 3d aplicadas al estudio de la cabeza del nervio óptico y enfermedades asociadas, 1/2017 - 12/2019. PR26/16-9B-1.

Rodríguez Sanjurjo, J. M. y Ruiz Sancho, J. M. *Introducción a la geometría diferencial. II: Superficies*. Sanz y Torres, 2018.

Rokach, L. y Maimon, O. Z. *Data mining with decision trees: theory and applications*, vol. 69. World scientific, 2008.

Roorda, A., Romero-Borja, F., Donnelly III, W. J., Queener, H., Hebert, T. J. y Campbell, M. C. Adaptive optics scanning laser ophthalmoscopy. *Optics express*, vol. 10(9), páginas 405–412, 2002.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, vol. 20, páginas 53–65, 1987.

Sadjadi, F. A. y Hall, E. L. Three-dimensional moment invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2), páginas 127–136, 1980.

Smith, L. I. A tutorial on principal components analysis. Informe técnico, 2002.

Sugumaran, V., Muralidharan, V. y Ramachandran, K. Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. *Mechanical systems and signal processing*, vol. 21(2), páginas 930–942, 2007.

Tortora, G. J. y Derrickson, B. *Principios de anatomía y fisiología*. Médica Panamericana„ 2013.

Upton, G. y Cook, I. *A dictionary of statistics 3e*. Oxford university press, 2014.

Wagner, S. y Wagner, D. *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

Xu, D. y Li, H. 3-d surface moment invariants. En *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, páginas 173–176. IEEE, 2006.

Xu, R. y Wunsch, D. *Clustering*, vol. 10. John Wiley & Sons, 2008.